Universidade de São Paulo
Instituto de Matemática e Estatística
Bacharelado em Ciência da Computação

Predição de Fluxos de Ciclistas Usuários de Sistemas de Compartilhamento de Bicicletas

Uma modelagem por aprendizado de máquina aplicada à Mobilidade Urbana e a Cidades Inteligentes

Éderson Cássio Lacerda Ferreira

Monografia Final

mac 0499 — Trabalho de Formatura Supervisionado

Supervisor: Prof. Dr. Fabio Kon
Cossupervisor: Prof. Dr. R. Hirata Jr.

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CNPq

São Paulo
20 de novembro de 2019

Resumo

Éderson Cássio Lacerda Ferreira. **Predição de Fluxos de Ciclistas Usuários de Sistemas de Compartilhamento de Bicicletas: *Uma modelagem por aprendizado de máquina aplicada à Mobilidade Urbana e a Cidades Inteligentes***. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

Sistemas de Compartilhamento de Bicicletas (SCB) têm surgido ao longo da última década como opção de mobilidade, em especial nos grandes centros urbanos, oferecidos tanto pela administração pública quanto pela iniciativa privada. Sua aceitação e efetivo uso pela população depende dos padrões de mobilidade na cidade e da existência de infraestrutura adequada para pedalar com segurança, na forma de ciclovias ou ciclofaixas separadas das vias para veículos motorizados, sinalização e outros.

Quando se deseja implantar tal sistema em uma cidade, uma pergunta surge: onde construir a infraestrutura para bicicletas? Para respondê-la, é preciso conhecer os locais de origens e destinos de viagens mais frequentes nas diversas horas do dia e épocas do ano, e quais dessas viagens são ou poderiam ser realizadas por bicicleta.

Este trabalho propõe uma abordagem para responder a essa pergunta através da modelagem computacional dos fluxos de ciclistas em um SCB existente, relacionando-os a variáveis geográficas, sociais, econômicas e meteorológicas dos locais e épocas em que as viagens ocorreram, gerando um modelo preditivo que pode ser extrapolado para outras regiões. Um modelo preditivo de regressão é desenvolvido a partir de uma abstração das viagens realizadas em regiões de origem e destino (fluxos), sendo os indicadores mencionados as variáveis de entrada, e a contagem de viagens realizadas dentro de cada fluxo a variável a ser predita pelo modelo.

São descritas em detalhes as fontes de dados usadas e a forma com que são integradas, bem como as limitações técnicas e computacionais encontradas.

Palavras-chave: Sistemas de Compartilhamento de Bicicletas. Origem e Destino. Mobilidade Urbana. Modelagem Preditiva. Aprendizado de Máquina. Regressão.

Abstract

Éderson Cássio Lacerda Ferreira. **Flow Prediction for Bike Sharing Cyclist Users: *A Machine Learning model applied to Urban Mobility and Smart Cities***. Undergraduate Thesis (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2019.

Bike Sharing Systems (BSS) have appeared over the last decade as a mobility option, mainly in large urban centers, offered by both public administration and private initiative. Acceptance and actual use of such system depends on mobility patterns inside the city and the existence of appropriate infrastructure for safely riding, in the form of cycle tracks or cycleways, separated from motor vehicles ways.

When someone wants to implement such a system in a city, a question arises: where should bike infrastructure be build? To answer it, it is needed to know the most performed paths by people in several hours of the day and periods of the year, and determine which of them, or which parts of them, are or could be performed by bicycle.

This work proposes an approach to answer that question through a computational modelling of cyclist flows in an existing BSS, connecting them to geographical, social, economical and meteorological variables of the places and time periods the trips occurred in, resulting in a predictive model that can be extrapolated to new places. It develops a predictive regression model using an abstraction of the performed trips in origin and destination regions (flows), being the mentioned indicators the input variables, and the in-flow trip counting the output variable to be predicted by the model.

The used data sources and the way they are integrated together, as well as the technical and computational limitations found are described in details.

Keywords: Bike Sharing Systems. Origin and Destination. Urban Mobility. Predictive Modelling. Machine Learning. Regression.

Sumário

Capítulo 1

Introdução

A implantação de Sistemas de Compartilhamento de Bicicletas nas cidades tem tido uma expansão ao longo da última década como alternativa de mobilidade para a população. Esse esforço visa mitigar os já há muito conhecidos problemas causados pelo excesso de carros nas vias urbanas, sendo os principais a poluição, os congestionamentos, e a falta de espaço para estacionamentos, bem como estimular a adoção de um hábito comprovadamente benéfico para a saúde física e mental.

Os sistemas de compartilhamento de bicicletas são serviços que as disponibilizam para aluguel ou empréstimo, oferecidos pelo poder público ou pela iniciativa privada. Podem apresentar-se na forma de sistemas com estações fixas, com locais determinados para retirada e devolução das bicicletas (não implicando que a bicicleta retirada em uma estação deva necessariamente ser devolvida nessa mesma estação), ou, mais recentemente, sem quaisquer estações (*dockless*), ficando as bicicletas disponíveis nas vias públicas e podendo até mesmo ser rastreadas e localizadas por GPS, através de aplicativo para smartphones. Na cidade de São Paulo, destacam-se os sistemas BikeSampa, CicloSampa (modelo com estações) e Yellow (sem estações). Na esteira da expansão desse tipo de sistema, novas empresas entram no mercado oferecendo, além de bicicletas, os patinetes elétricos.

No entanto, o deslocamento por bicicleta em vias de tráfego intenso requer a existência de infraestrutura adequada. Em São Paulo, a nomenclatura diferencia as *ciclovias*, vias segregadas para bicicletas, e as *ciclofaixas*, porções das vias demarcadas apenas por pintura, podendo haver horários em que estão de fato disponíveis para uso exclusivo dos ciclistas. Outras cidades ou países adotam nomenclaturas de alguma forma semelhantes, levando em consideração a segregação ou compartilhamento do espaço com o tráfego de veículos ou de pedestres.

[1] https://tembici.com.br/bicicletas-compartilhadas. Acessado em: 24 de outubro de 2019.

Figura 1.1: *Estação fixa do sistema BikeSampa, operado pela empresa Tembici. Fonte: site da Tembici*[1]

Determinar locais para a construção de uma infraestrutura desse tipo em uma cidade requer o conhecimento dos padrões de mobilidade da sua população. Bicicletas são usadas principalmente em viagens de curta distância e duração, ou combinadas com o transporte público, como solução para o problema da "primeira e última milha". A mobilidade pode ser influenciada pelos indicadores sociais e econômicos das diferentes regiões da cidade, pela existência de áreas comerciais ou com muitos postos de trabalho, pelas condições climáticas do dia ou estação do ano, e muitos outros fatores. É muito difícil quantificar quantas variáveis estão relacionadas à mobilidade e, em especial, ao uso de bicicletas em determinados deslocamentos.

Os padrões de mobilidade são modelados geralmente como pares de regiões de origem e de destino, aqui chamados *fluxos*. Dadas duas regiões A e B do espaço geográfico da cidade, são quantificadas as viagens feitas por bicicleta que partem de A e chegam em B em determinado período de tempo. O *fluxo* é a unidade básica de modelagem, definida em detalhes na seção 3.1.

A proposta deste trabalho é desenvolver um modelo dos fluxos, obtido por aprendi-

zado de máquina, a partir de dados reais de Sistemas de Compartilhamento de Bicicletas, relacionando-os a diversos indicadores que caracterizam os locais de origem e destino das viagens, bem como a situação meteorológica nos períodos de tempo considerados, de tal forma que a caracterização e predição possam ser feitas para diferentes áreas onde o sistema ainda não esteja implantado.

Em particular, pretende-se tentar responder às seguintes questões de pesquisa:

1. É possível criar um modelo preditivo de quais deslocamentos de bicicletas são mais importantes (pares de regiões de origem e destino) e aplicá-lo em uma dada região urbana?

2. Quais as variáveis mais importantes que influenciam os padrões de mobilidade em Sistemas de Compartilhamento de Bicicletas?

O trabalho é realizado no contexto do projeto *BikeScience*, braço do grupo de pesquisa em Cidades Inteligentes *InterSCity* (http://interscity.org/). O BikeScience está desenvolvendo trabalho acadêmico sobre abstração de fluxos de mobilidade em Sistemas de Compartilhamento de Bicicletas, tendo-os estudado de maneira descritiva, e a próxima etapa é a modelagem e predição desses fluxos por aprendizado de máquina.

Os dados usados neste trabalho foram obtidos a partir das viagens do Bluebikes, um sistema com estações fixas da cidade de Boston, nos Estados Unidos, e do Indego, da cidade da Filadélfia, no mesmo país. O período abarcado é de um (1) ano, de abril de 2018 a março de 2019, como forma de capturar a influência das estações do ano no uso de bicicletas pela população.

As principais variáveis de caracterização do modelo foram obtidas de outras fontes de dados que não as dos Sistemas de Compartilhamento de Bicicletas. Entre as fontes de dados externas aos Sistemas, foram escolhidas:

- Dados censitários. As agências governamentais proveem a divisão da área de cobertura em diversos níveis de granularidade, normalmente trazendo setores do tamanho de alguns quarteirões. Cada setor tem seus indicadores populacionais organizados por sexo, idade, renda, nível de escolaridade e outras variáveis.

- Dados climáticos históricos. Existem diversos serviços em nuvem (*cloud*) que proveem a captura de indicadores meteorológicos por período de tempo e localidade. Entre as variáveis disponíveis, estão temperatura, umidade, velocidade e direção do vento, visibilidade e outras.

- Pontos de interesse. A existência de pontos comerciais, institucionais, transporte público, áreas de lazer ou trabalho em diferentes ramos concentrada em diversas

regiões e dispersa ou ausente em outras, foi modelada de forma a corresponder a concentração de pontos de interesse de diversos tipos com as regiões de origem ou destino das viagens feitas por bicicleta.

- Elevação. A existência de ladeiras em uma cidade pode determinar a viabilidade de viajar por bicicleta entre dois pontos A e B, assim as elevações ou altitudes das regiões de origem e destino também entram como variáveis no modelo.
- Estrutura cicloviária: o histórico de viagens é altamente influenciado pela existência de ciclovias e ciclofaixas, criando um viés em favor de áreas em que essa infraestrutura é existente. É preciso que o modelo capture esse viés.

O trabalho começa, no capítulo 2, descrevendo o arcabouço tecnológico utilizado. O aprendizado de máquina é apresentado conceitualmente e as ferramentas computacionais são introduzidas. A linguagem de programação Python apresenta um consagrado ecossistema de análise de dados, com bibliotecas para processamento de conjuntos de dados em formato de tabela (*Pandas*), processamento geoespacial (*GeoPandas*), plotagem de gráficos (*Matplotlib*) e mapas (*Folium*), bem como de aprendizado de máquina (*Scikit-Learn*).

O capítulo 3 discute como é criado um modelo de fluxos de mobilidade (também conhecidos como fluxos de *origem-destino*), a forma como esses fluxos são computados, agregados e plotados em mapas, e como são determinados os fluxos mais relevantes. São descritos os conjuntos de dados externos que agregam caracterização aos fluxos e como são integrados em um único conjunto de amostragem para o algoritmo de aprendizado. Por fim, o *pipeline* de processamento de aprendizado de máquina é discutido, apresentando como o modelo é gerado, testado, validado e refinado.

Antes da conclusão do trabalho, o capítulo 5 descreve os resultados obtidos e procura comparar sua qualidade com os dados do Bluebikes e do Indego, levando em conta as características de cada sistema e dos conjuntos de dados disponíveis.

Capítulo 2

Conceitos fundamentais

Este capítulo destina-se a descrever o arcabouço tecnológico sobre o qual o trabalho foi desenvolvido. O desenvolvimento tecnológico recente está permitindo às organizações rastrear e coletar dados sobre tudo o que os usuários de sistemas computacionais fazem. A criação computacional de modelos preditivos requer massivo processamento de dados e o uso de *aprendizado de máquina*.

O desenvolvimento da assim chamada *internet das coisas* leva a conectividade a diversos utensílios do dia a dia, aumentando o volume de dados gerados. Neste contexto encaixam-se os Sistemas de Compartilhamento de Bicicletas, onde os usuários podem realizar as transações de retirada e devolução através de equipamentos instalados em totens nas estações (caso do Bluebikes e do Indego) ou diretamente na bicicleta (como exemplo, a Yellow de São Paulo), sem intermediários, no máximo com auxílio do smartphone.

Quando se coleta quaisquer tipo de dados, em grande volume, uma pergunta que surge é: como aproveitá-los? Como podem ser úteis?

2.1 Aprendizado de máquina

Tradicionalmente, a abordagem computacional para resolver problemas envolve a descrição precisa de algoritmos e fórmulas matemáticas, entregando soluções fechadas para problemas que as admitem. Porém, problemas sem resposta analítica podem ter suas respostas ao menos aproximadas pelo computador? Por exemplo, não existe uma definição matemática de “árvore”, porém uma criança de 3 anos sabe reconhecer uma não porque aprendeu uma definição, e sim por ter visto algumas poucas árvores. Seria o computador capaz de uma façanha semelhante?

Segundo Yaser S. Abu-Mostafa (2012), "aprendizado a partir de dados é usado em situações onde não temos uma solução analítica, mas temos dados que podem ser usados para construir uma solução empírica". Os autores definem formalmente o *problema do aprendizado* da seguinte forma:

- existe um espaço X dos possíveis dados de entrada e um espaço Y das possíveis respostas para o problema
- existe uma *função alvo* $f : X \longrightarrow Y$, desconhecida
- há um conjunto de dados de exemplo $D = \{(x, y)\}, x \in X, y \in Y$, de forma que $f(x) = y$
- deseja-se encontrar uma função $g : X \longrightarrow Y$ que aproxime f

Um *algoritmo de aprendizado* é um algoritmo que, a partir de um conjunto de dados D e um *espaço de hipóteses* H, determina uma função $g \in H$ que, espera-se, aproxima f. O espaço de hipóteses é um conjunto de funções candidatas de natureza semelhante, a depender do algoritmo usado. Por exemplo, no caso de uma *regressão polinomial* (Figura 2.1), tenta-se determinar um polinômio de algum grau n dado que melhor se ajuste aos dados, isto é, quais seriam os coeficientes que multiplicariam os atributos dos elementos x de dados. Já uma *árvore de decisão* (Figura 2.2) é uma sequência de decisões sobre os atributos e suas faixas de valores, a qual pode ser representada em código como um conjunto de condicionais encadeados. O algoritmo de aprendizado monta a árvore procurando quais atributos melhor separam os dados e quais valores servem como pontos de separação, para que ao final de uma execução dos condicionais se tenha uma predição.

2.1.1 Tipos de aprendizado de máquina

Quando as amostras de dados trazem as saídas ou respostas corretas para esses dados de exemplo, isso caracteriza o *aprendizado supervisionado*. Também existem formas de *aprendizado não supervisionado*, as quais se destinam a encontrar padrões nos dados, sem que se espere de antemão qual resposta deve ser encontrada. Um exemplo muito usado de aprendizado não supervisionado é a *clusterização*, que é a separação dos elementos de dados em conjuntos com características semelhantes entre si e diversas dos outros conjuntos. É usada na segmentação de clientes ou usuários de empresas, como forma de extrair perfis de pessoas e direcionar anúncios publicitários mais efetivos para cada perfil.

[1] https://www.r-bloggers.com/fitting-polynomial-regression-in-r/. Acessado em: 24 de outubro de 2019.
[2] https://www.cs.cmu.edu/ bhiksha/courses/10-601/decisiontrees/. Acessado em: 24 de outubro de 2019.

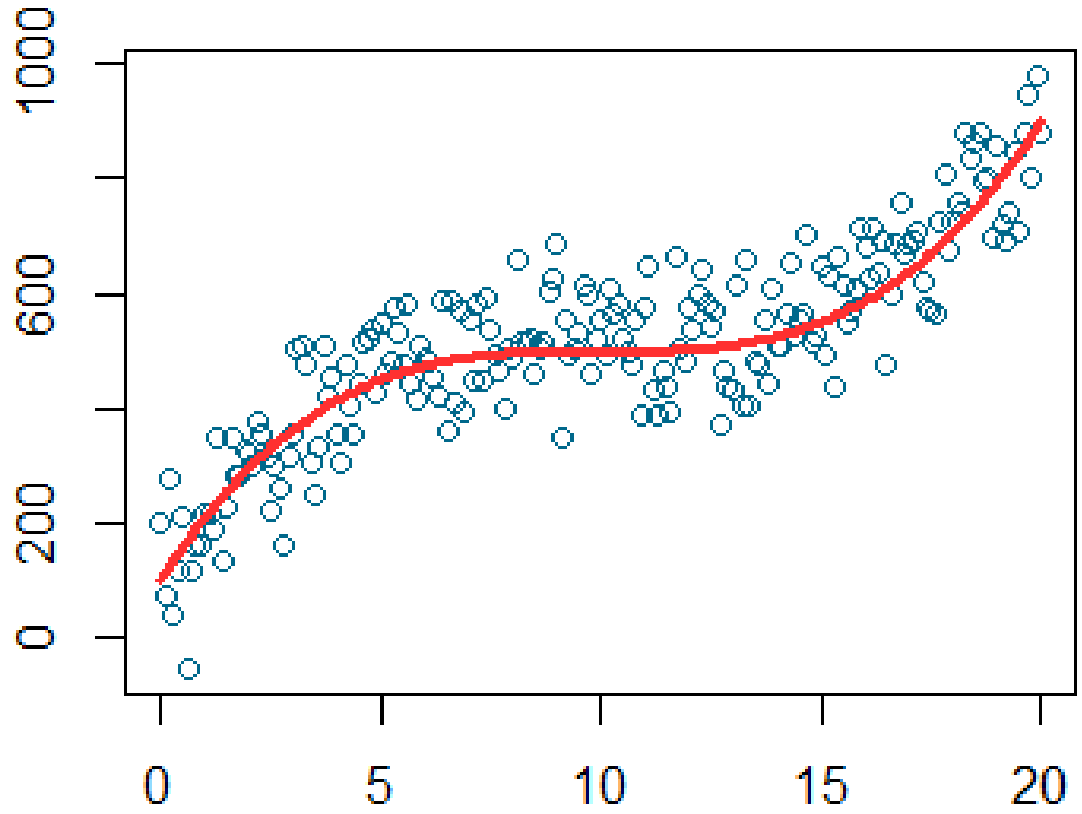

Figura 2.1: *Regressão polinomial: a procura de uma função polinomial (linha vermelha) que se ajusta às amostras de dados (circunferências). O eixo horizontal é um atributo dos dados (regressão com uma única variável) e o vertical é a variável alvo de predição. Fonte: site R-bloggers*[1]

Os problemas onde o aprendizado supervisionado é aplicado podem ser de classificação ou de regressão. Um problema de *classificação* consiste em atribuir um rótulo entre vários definidos, por exemplo: diagnosticar uma doença a partir da imagem de uma tomografia (uma *classificação binária*, a emissão de um rótulo do tipo "sim" ou "não"), aprovar ou não um crédito bancário para um cliente a partir de seu histórico, encaixar qualquer objeto em uma de duas ou mais categorias. Um problema de *regressão* consiste em atribuir um valor numérico a uma grandeza, como se houvesse uma função matemática que a calcularia. Por exemplo, atribuir preços a imóveis, estimar projeções de lucros ou crescimento econômico, estimar quantas viagens entre duas regiões da cidade os usuários de um Sistema de Compartilhamento de Bicicletas farão ao longo de um mês. Em ambos os tipos de aprendizado supervisionado, objetos pré-classificados ou com valor numérico já atribuído são fornecidos ao algoritmo de aprendizado como entrada para a construção de um modelo preditivo.

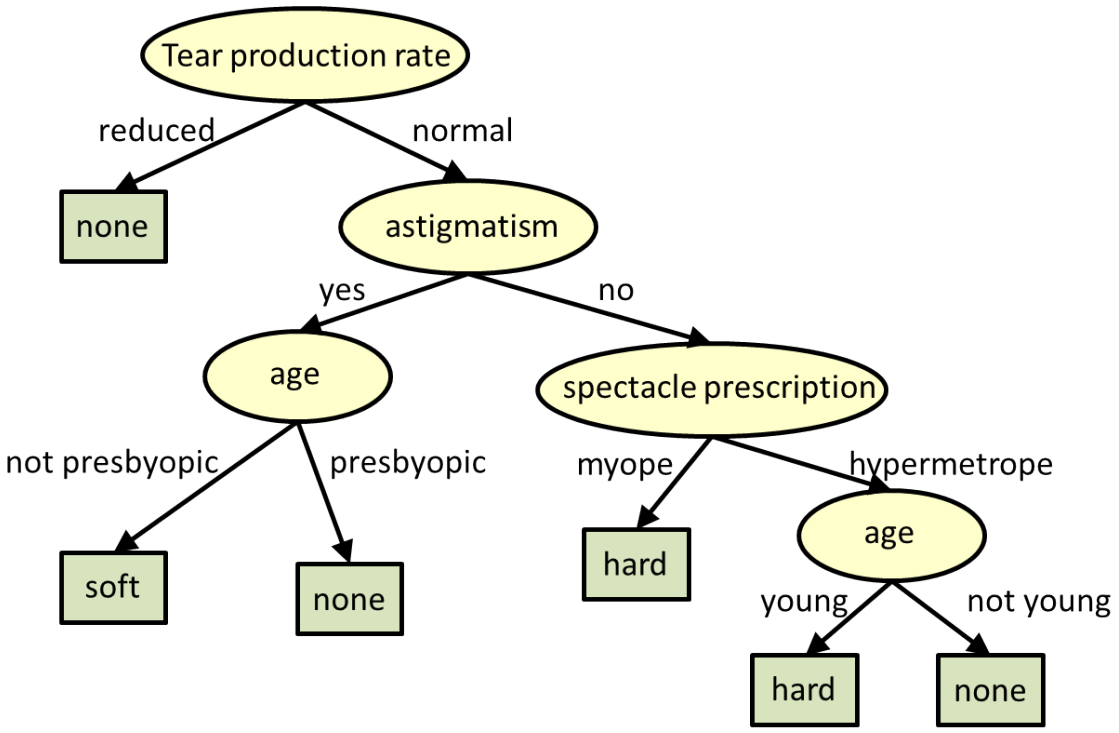

Figura 2.2: *Árvore de decisão Os ovais representam condicionais sobre os atributos de dados, e os retângulos são as predições emitidas pelo modelo. Fonte: site do professor Bhiksha Raj do Carnegie Mellon School of Computer Science*[2]

2.1.2 *Pipeline*: preparação, ajuste, validação e teste

Executar uma tarefa de aprendizado envolve uma série de transformações nos dados e, algumas vezes elas podem ser padronizadas em um encadeamento (ou *pipeline*), que é um conjunto de ações encadeadas necessárias para construir um modelo preditivo a partir de dados através de aprendizado de máquina.

Normalmente temos à disposição dados brutos, os quais não podem ser fornecidos diretamente como entrada para os algoritmos de aprendizado pois podem possuir partes incompletas e conter erros ou ruído que não refletem a realidade sendo modelada. Uma primeira etapa, então, é garantir a limpeza dos dados, eliminando os elementos de dados com problemas ou substituindo os valores espúrios pela média ou mediana do conjunto.

Por exemplo, um algoritmo de regressão do tipo linear ou polinomial, que trabalha minimizando uma função de erro, não aceitaria atributos contendo, por exemplo, valores *string* como “Masculino” ou “Feminino”. Uma primeira solução seria codificar as categorias como números, o que implicaria em atribuir-lhes um valor de grandeza que pode introduzir vieses no algoritmo. Muitas vezes o recomendado é remover o atributo categórico e criar um atributo para cada possível categoria, os quais terão valores 0 ou 1 a depender da categorização do elemento. Esses atributos são chamados de *dummy attributes*.

Outra operação importante de preparação dos dados é a *"normalização"*. Alguns algoritmos podem ser sensíveis à escala em que estão os dados, ou seja, atributos com altos valores absolutos podem ter um peso maior e influir no aprendizado. A normalização ajusta os valores de todos os atributos para uma mesma escala, por exemplo, entre 0 e 1, ou calculando uma estatística como a *estatística Z* (distribuição normal padrão), onde a média do atributo passa a valer zero e os outros valores passam a ser medidos pela distância da média em desvios padrões.

A execução do aprendizado é chamada, algumas vezes, de *ajuste*. Isto é, um modelo é iniciado com parâmetros aleatórios e vai se aproximando dos valores ótimos de acordo com alguma medida de perda, conforme o algoritmo lê os dados e executa algum processamento sobre eles. Como exemplo, considere-se uma regressão do tipo polinomial, isto é, ajustar um polinômio de grau n a pontos de dados em um espaço X. Os parâmetros do modelo são os coeficientes do polinômio. Uma *função de erro* é usada para determinar a qualidade do ajuste – podendo, neste caso, ser o erro quadrático de uma predição, isto é, o quadrado da diferença entre o valor real de um elemento x da amostra e o valor atualmente predito. A função de erro é uma função dos parâmetros, e não dos elementos de dados, considerados constantes. O algoritmo *gradiente descendente* pode ser usado para minimizar uma função de erro "caminhando" pelo espaço dos parâmetros seguindo o sentido contrário ao vetor gradiente (Yaser S. Abu-Mostafa, 2012).

O próprio algoritmo de aprendizado pode ser parametrizado (*hiperparâmetros*) e ser otimizado na etapa de *validação*. Essa etapa consiste em procurar valores adequados de hiperparâmetros. Uma maneira simples de realizar a validação é particionar os dados da amostra em dois conjuntos (treino e validação) e efetuar ajustes para cada combinação de hiperparâmetros que se desejar, ficando com aquela que oferecer a maior acurácia sobre a porção reservada. A técnica de *validação cruzada* consiste em particionar o conjunto de dados em n porções e, a cada vez, fazer o ajuste usando $n - 1$ partes para treino e uma para validação. O erro considerado é a média dos erros para cada porção. É uma técnica custosa em tempo de processamento, pois todas as combinações desejadas de hiperparâmetros (que podem ser muitas) têm de ser ajustadas n vezes. Por outro lado, usar somente uma porção fixa dos dados para validação pode enviesar o modelo final.

Por último, a etapa de *teste*, que consiste em dar o veredito final sobre a acurácia do modelo a partir de um conjunto de dados guardado *somente para esse fim*. Essa etapa não pode ser confundida com a de validação, onde também se separa dados para validar os hiperparâmetros do algoritmo de aprendizado. Os dados de validação são revisitados inúmeras vezes no processo e decisões de hiperparâmetros são tomadas com base neles. O objetivo do conjunto de dados de teste é verificar e possivelmente atestar a capacidade de

generalização do modelo para além dos dados usados no aprendizado.

2.1.3 Overfitting

Um problema comum quando se lida com aprendizado de máquina é quando o modelo aprendido se ajusta tão perfeitamente aos dados de treinamento e não acerta bem nos dados de teste (ou no mundo real). Ou seja, o modelo não *"generaliza"* bem. Nesses casos dizemos que o modelo "decorou" os dados, analogamente a um estudante que decorou a matéria sem a entender e foi mal em uma prova. Esse problema é denominado *overfitting* (Figura 2.3).

A principal causa do *overfitting* é a existência de um componente aleatório nos dados, o qual gera ruído na função alvo sendo aprendida (Yaser S. Abu-Mostafa, 2012). Por exemplo, duas pessoas com características (quase) idênticas podem dar notas muito diferentes para um mesmo vídeo do YouTube, assim recomendar determinado vídeo para uma não surtirá o mesmo efeito que para outra.

A etapa de *validação*, além de selecionar hiperparâmetros, busca evitar o *overfitting* ao se efetuar o ajuste de um modelo com somente uma parte do conjunto de dados. Outra técnica possível é a *regularização*, que restringe o espaço de busca do algoritmo de forma a evitar encaixes muito perfeitos dos parâmetros aos dados.

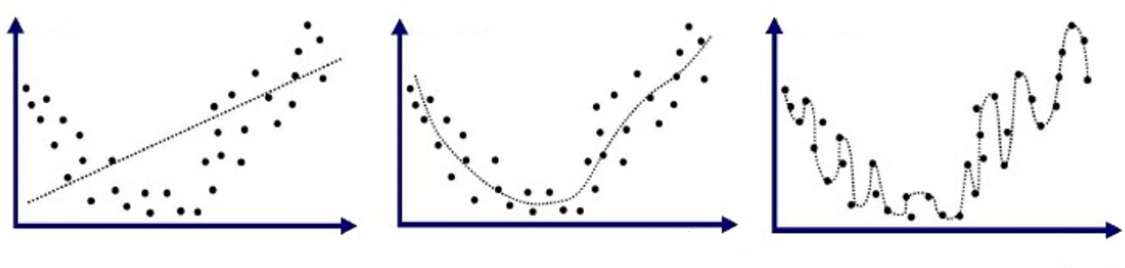

Figura 2.3: *Exemplos de modelos (linhas pontinhadas) para regressão polinomial a partir de amostras (pontos). À esquerda, o encaixe de um modelo linear apresenta pouco ajuste (underfitting). No meio, um modelo considerado robusto. À direita, um modelo sobreencaixado (overfitting). Fonte: artigo de Anup Bhande no site Medium*[3]

2.2 Tecnologias

A linguagem de programação Python é uma escolha comum quando se fala em análise de dados e aprendizado de máquina, devido à disponibilidade de bibliotecas de código

[3] https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76. Acessado em: 24 de outubro de 2019.

aberto. Esta seção descreve algumas das bibliotecas do Python utilizadas no processamento de dados e na geração de modelos a partir de dados de Sistemas de Compartilhamento de Bicicletas e outras fontes.

2.2.1 Jupyter Notebook

O Jupyter Notebook (https://jupyter.org/) é uma aplicação Web que permite a criação e o compartilhamento de documentos que mesclam texto, código Python e a sua saída textual ou gráfica (Figura 2.4), também conhecidos como "cadernos". É dito ser uma ferramenta de *storytelling*, ou contagem de histórias (Sampaio, 2018), pois todas as etapas de processamento e análise de dados podem ser descritas e apresentadas em uma ordem coerente, com seu código e sua saída.

Uma característica interessante do Jupyter é que o código pode ser continuamente reexecutado, alterando o conteúdo do caderno ao qual pertence. Assim, os cadernos permitem a reprodutibilidade das análises, bastando ter os dados de entrada à disposição.

Todo o código para este trabalho foi desenvolvido em uma sequência de cadernos que descreve todos os passos de análise e processamento realizados. Cada caderno salva seus resultados intermediários para que possam ser lidos como entrada nos próximos.

2.2.2 Pandas

A biblioteca Pandas (https://pandas.pydata.org/) provê estruturas de dados e operações de alto desempenho sobre conjuntos de dados em formato tabular, chamados *data frames* , que funcionam como um sistema gerenciador de banco de dados relacional, porém operando em memória.

Um *data frame* é um conjunto de colunas (*data series*) de dados que compartilham um índice (*index*) comum (Figura 2.5). O índice identifica ou rotula cada elemento de dados em uma coluna, e em um *data frame* ele identifica as linhas de dados, enquanto as colunas representam os atributos de cada linha. Por default, o índice é uma numeração sequencial, mas qualquer subconjunto das colunas pode ser transformada no índice do *data frame*.

Dentre as operações oferecidas pelo Pandas, vale destacar (Community, 2019):

- Tratamento de dados faltantes
- Inserção e exclusão de colunas
- Alinhamento dos itens de dados nas colunas conforme o índice comum: itens em

Gender distribution

In [6]:
```
pd.crosstab( trips["usertype"], trips["gender"], normalize='index', margins=True)*100
```

Out[6]:

gender	0.0	1.0	2.0
usertype			
Customer	62.4	28.3	9.4
Subscriber	5.8	70.5	23.7
All	15.3	63.4	21.3

Trip duration distribution

In [7]:
```
duration = trips[trips['tripduration'] < 3000]

duration = duration[['tripduration']]

ax_duration = plt.axes()
ax_duration.set_axisbelow(True)
plt.grid(linestyle='--')
ax_duration.xaxis.grid(False)
formatter = tkr.FuncFormatter(ch.numbers_in_thousands)
ax_duration.yaxis.set_major_formatter(formatter)
plt.hist(bins=100,x=duration['tripduration']/60)
plt.title('Duration')
plt.xlabel('Minutes')
plt.ylabel('Trips (in thousands)')
fig_duration = plt.gcf()
```

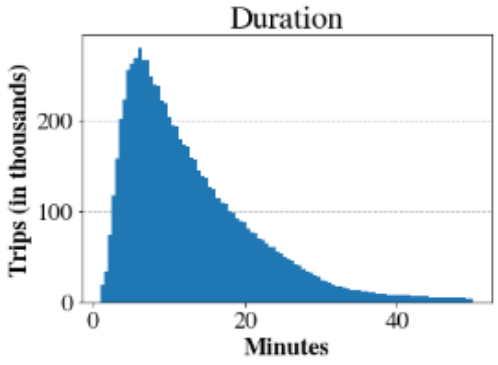

Figura 2.4: *Documento criado com o Jupyter Notebook. Fonte: elaborado pelo autor*

diferentes colunas com o mesmo índice pertencerão à mesma linha

- Agrupamento dos dados por conjuntos de colunas, e operações de agregaçao sobre os grupos (soma, média, contagem, desvio padrão e outras) (Figura 2.6)
- Conversão de objetos como listas Python, vetores NumPy em estruturas de dados do Pandas
- Fatiamento e amostragem
- Mesclagem de diferentes *dataframes* por colunas comuns (junção) (Figura **fig:pandas-merging**)
- Rotulação multivalorada (hierárquica) das linhas e colunas
- Reformatação de um índice multivalorado em matriz e vice-versa (pivoteamento)
- Entrada e saída para diversos formatos de dados, como CSV, Excel, bancos de dados
- Tratamento de séries temporais

data series

index

	blunt_phrase	class	clds	day_ind	dewPt	expire_time_gmt	feels_like	gust	heat_index	icon_extd
0	None	observation	OVC	N	-2.0	1522565640	1.0	NaN	5.0	2600
1	None	observation	OVC	N	-2.0	1522569240	1.0	NaN	5.0	2600
2	None	observation	OVC	N	-2.0	1522572840	2.0	NaN	6.0	2600
3	None	observation	OVC	N	-2.0	1522576440	2.0	44.0	6.0	2600
4	None	observation	OVC	N	-2.0	1522580040	2.0	44.0	7.0	2690

Figura 2.5: *Data frame, a principal estrutura de dados do Pandas. Fonte: elaborado pelo autor*

2.2.3 GeoPandas

O GeoPandas (http://geopandas.org/) estende o Pandas, adicionando-lhe capacidades de *processamento geoespacial*. São comuns os conjuntos de dados com informações geográficas, onde os objetos consistem em pontos (ex.: estações de um Sistema de Compartilhamento de Bicicletas, pontos de ônibus, origens e destinos de viagens), linhas (ex.: ruas, ciclovias, trajetórias de viagens) e áreas poligonais (ex.: países, estados, municípios, bairros).

Além dos atributos de dados, os conjuntos de dados geoespaciais necessariamente apresentam um atributo *geometry*, que no GeoPandas é representado por um conjunto de objetos da biblioteca *Shapely* do Python. A Figura 2.8 mostra um exemplo de uso do GeoPandas, carregando um arquivo com a estrutura cicloviária de São Paulo, exibindo seu *data frame* e apresentando os objetos geométricos (ciclovias).

```
In [19]: #trips
trips_by_day = trips.groupby(['end station name'], as_index=False).agg({'tripduration': 'count'})
trips_by_day.columns = ['end station name', 'trip count']
trips_by_day.set_index('end station name',inplace=True)
trips_by_day.sort_values('trip count', ascending=False).head(10)
```

Out[19]:

end station name	trip count
MIT at Mass Ave / Amherst St	255127
MIT Stata Center at Vassar St / Main St	202514
Central Square at Mass Ave / Essex St	174599
South Station - 700 Atlantic Ave	171469
Harvard Square at Mass Ave/ Dunster	168416
Ames St at Main St	124915
Copley Square - Dartmouth St at Boylston St	124764
One Kendall Square at Hampshire St / Portland St	106852
Lafayette Square at Mass Ave / Main St / Columbia St	103052
Beacon St at Massachusetts Ave	101920

Figura 2.6: *Exemplo de agregação com Pandas: contando quantas viagens chegaram em cada estação do sistema Bluebikes, de Boston, e exibindo em ordem decrescente (ranking). Fonte: elaborado pelo autor*

Uma importante operação provida pela biblioteca é a *junção espacial*. Uma junção (ou mesclagem, para usar a terminologia do Pandas) é uma operação que confronta os elementos de dados de dois conjuntos, selecionando os pares de elementos que apresentam um ou mais atributos em comum (Figura 2.7). A junção espacial confronta objetos geoespaciais, selecionando os pares que se intersectam ou que estão contidos um em outro (por exemplo, determinar quais estações de bicicletas se encontram em determinado quarteirão, ou que viagens passaram por quais ruas).

Além da junção espacial, outras operações são oferecidas:

- Obter relações geométricas entre os elementos (intersecta, contém, está contido)
- Operações sobre elementos individuais providas pela biblioteca *Shapely*, como centroide, envelope, pontos equidistantes (*buffering*), transformações afins
- Agregação geométrica, combinando elementos com atributos de dados comuns
- Transformação entre sistemas de coordenadas ou *projeções*, como Mercator, azimutal e outras

2.2.4 Scikit-learn

Scikit-learn (https://scikit-learn.org/stable/) é a tradicional biblioteca de aprendizado de máquina do ecossistema Python, trazendo um conjunto diversificado de algoritmos para problemas de aprendizado supervisionado e não supervisionado, bem como ferramentas

```
In [17]: trip_counts.head()
```

Out[17]:

	hour	trip_count
0	2019-01-01 00:00:00	7
1	2019-01-01 01:00:00	24
2	2019-01-01 02:00:00	5
3	2019-01-01 03:00:00	3
4	2019-01-01 04:00:00	1

```
In [18]: mean_temperatures.head()
```

Out[18]:

	hour	mean_temperature
0	2018-04-01 00:00:00	5.0
1	2018-04-01 01:00:00	5.0
2	2018-04-01 02:00:00	6.0
3	2018-04-01 03:00:00	6.0
4	2018-04-01 04:00:00	7.0

```
In [19]: merge = trip_counts.merge(mean_temperatures, on='hour')
         merge.head()
```

Out[19]:

	hour	trip_count	mean_temperature
0	2019-01-01 00:00:00	7	4.0
1	2019-01-01 01:00:00	24	6.0
2	2019-01-01 02:00:00	5	6.0
3	2019-01-01 03:00:00	3	8.0
4	2019-01-01 04:00:00	1	8.0

Figura 2.7: *Exemplo de mesclagem com Pandas: dadas contagens de viagens por hora, e temperaturas médias por hora, juntar os dois dataframes pela coluna comum. Fonte: elaborado pelo autor*

```
In [7]: import geopandas as gpd
bike_sirgas = gpd.read_file('../data/sao-paulo/geosampa/SIRGAS_SHP_redecicloviaria/SIRGAS_SHP_redecicloviari
bike_sirgas.crs = {'init': 'epsg:31983'}
bike_sirgas.to_crs({'init': 'epsg:4326'}, inplace=True)
bike_sirgas.head()
```

Out[7]:

	rc_nome	rc_extensa	rc_inaugur	rc_descric	rc_tipo	rc_id	geometry
0	CICLOFAIXA VISCONDE DE TAUNAY	110	2015-01-23	CICLOFAIXA	ciclofaixa	10190	(LINESTRING (-46.71166104 -23.6365839899999, -...
1	CICLOFAIXA JAIR RIBEIRO	70	2016-06-23	CICLOFAIXA	ciclofaixa	10204	(LINESTRING (-46.68596100000002 -23.7034430099...
2	CICLOFAIXA INAJAR DE SOUZA - TRECHO IV	486	2016-10-03	CICLOFAIXA	ciclofaixa	10240	(LINESTRING (-46.67480399999994 -23.4547779999...
3	CICLOFAIXA CENTRO - ETAPA 1	105	2014-06-07	CICLOFAIXA	ciclofaixa	10589	(LINESTRING (-46.63917683999996 -23.5349710199...
4	CICLOFAIXA JOSE ALVES CUNHA LIMA	434	2015-02-25	CICLOFAIXA	ciclofaixa	10315	(LINESTRING (-46.74688415999998 -23.5643979599...

```
In [11]: import matplotlib.pyplot as plt
fig, ax = plt.subplots(figsize=(7, 7))
bike_sirgas.plot(ax=ax)
plt.show()
```

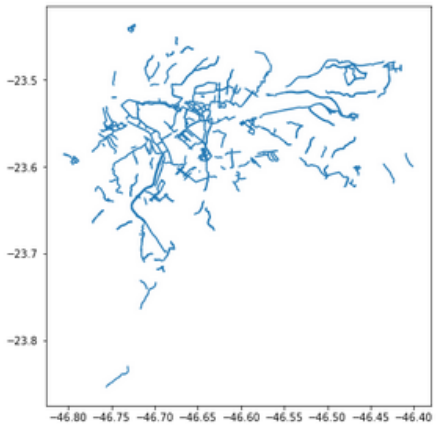

Figura 2.8: *Uso do GeoPandas: carga e exibição de um arquivo de informações geoespaciais. Fonte: elaborado pelo autor*

para o tratamento e preparação dos dados e composição do *pipeline* (encadeamento) das operações.

A biblioteca traz um vasto conjunto de ferramentas, entre as quais destacam-se:

- Algoritmos para regressão, classificação, clusterização
- Seleção de modelos: grid de hiperparâmetros, validação cruzada, busca de hiperparâmetros sem força bruta, funções de erro e métricas de acurácia para regressão e classificação
- Seleção de atributos: análise de componentes principais, análise de componentes independentes e outros
- Preprocessamento: normalização, transfomações não lineares, codificação de variáveis categóricas, discretização
- Pipeline: composição, persistência

O recurso de *pipeline* permite encadear a sequência de operações necessárias para uma

tarefa de aprendizado de máquina (subseção 2.1.2). Um pipeline montado pode ter seus parâmetros ajustados e persistido com o auxílio da biblioteca *Joblib* para ser reaplicado em predições.

2.2.5 Outras

Outras bibliotecas Python consagradas usadas neste trabalho foram:

- **Matplotlib** (https://matplotlib.org/): é a tradicional biblioteca Python para criação de gráficos. Todos os gráficos deste trabalho foram criados com esta ferramenta.
- **Folium** (https://python-visualization.github.io/folium/): para a geração de mapas. Integra a biblioteca JavaScript *Leaflet* (https://leafletjs.com/), a qual permite criar mapas interativos para aplicações web e móveis, com cadernos do Jupyter.
- **Shapely** (https://shapely.readthedocs.io/en/stable/manual.html): realiza manipulação de objetos geométricos.
- **Geopy** (https://geopy.readthedocs.io/en/stable/): manipula coordenadas geográficas.
- **Holidays** (https://pypi.org/project/holidays/): contém datas de feriados em diversos países; útil para a caracterização temporal.
- **Haversine** (https://pypi.org/project/haversine/): método matemático para estimativa de distâncias sobre superfícies esféricas. Calcula a distância entre duas coordenadas geográficas de maneira mais precisa do que com uma distância euclidiana.
- **OSMnx** (https://osmnx.readthedocs.io/en/stable/): acessa dados viários do OpenStreetMap e pode processá-los como grafos.
- **Joblib** (https://joblib.readthedocs.io/en/latest/): persiste uma sequência de transformações executadas em dados e permite reaplicá-las em novos dados.

Capítulo 3

Modelo de aprendizado de máquina para fluxos de bicicletas compartilhadas

3.1 Fluxos de origem e destino

Para entender padrões de mobilidade de uma população em uma determinada área e de uso de determinado meio de transporte, são muito comuns análises e pesquisas denominadas *origem-destino*. Este tipo de análise pretende determinar pontos ou subregiões de onde partem ou para onde vão as pessoas em suas viagens em seu dia-a-dia, para trabalho, estudo, lazer ou qualquer atividade. Certos movimentos de *comutação pendular* são notórios, com grande quantidade de pessoas deslocando-se de regiões predominantemente residenciais para regiões predominantemente comerciais ou industriais durante a manhã, e fazendo o caminho inverso ao final da tarde.

Porém, conhecer em detalhes os padrões de mobilidade requer extensiva análise das viagens realizadas, para as quais não existem dados coletados ou integrados em se tratando de todas as viagens realizadas pela população em geral. Dessa forma, pesquisas em origem-destino como a realizada por Metropolitano de São Paulo (2019) baseiam-se em entrevistas com uma amostra da população.

Este trabalho pretende analisar os padrões de origem e destino em um contexto mais específico, o dos Sistemas de Compartilhamento de Bicicletas. Dispondo-se dos dados de instante, ponto de origem e ponto de destino de cada viagem realizada através do sistema, o comportamento de seus usuários nesse contexto pode ser modelado sem preocupações com o uso de outros modais de transporte – embora, como será discutido adiante, a

disponibilidade de serviços de ônibus, metrô e outros relaciona-se de alguma maneira com o fluxo de ciclistas em determinadas regiões.

3.1.1 Definição de fluxo

Nesta seção será definido o conceito de *fluxo* como o principal objeto de modelagem. A ideia é modelar a quantidade de viagens que partem de uma região A e chegam em uma região B em determinado período do tempo. Para isso, precisamos definir formalmente as regiões geográficas e o que é exatamente um período no tempo. Por exemplo, perguntar "qual o fluxo de viagens de bicicleta entre duas estações de metrô no período da manhã?" pode dar uma boa ideia intuitiva, mas ainda é vaga para fins de modelagem.

Primeiro, como delimitar quais viagens partem das referidas estações de metrô? Caso a pessoa, ao sair de uma estação, tenha necessitado andar 10 metros para ter acesso ao Sistema de Compartilhamento de Bicicletas, é considerado que ela partiu da estação? E se andou, 50, 100 metros? Analogamente, o que significa "manhã"? Claro, podemos usar o conceito do senso comum, mas se a análise descritiva realizada nos dados (ver Figuras 3.6 e 3.7) revelar horários de pico, podemos considerar toda a manhã como sendo o mesmo tipo de período do dia?

Os exemplos acima são apenas algumas reflexões para elucidar que é preciso ter uma delimitação precisa do que se chama de regiões do espaço geográfico e de períodos do tempo. Começando pelo espaço, o exemplo do deslocamento entre dois pontos bem específicos (duas estações de metrô) sugere uma divisão de alta granularidade da área considerada, mas perguntas do tipo também podem ser feitas para quarteirões, bairros ou regiões administrativas maiores de uma cidade. Quanto ao tempo, o que se quer considerar, uma manhã específica ou todas as manhãs de maneira geral (pelo menos as de dias úteis)? Se os dados revelam padrões diferentes para diferentes meses do ano, períodos do dia, dias úteis ou fins de semana e feriados, essa separação pode ser útil para a modelagem.

Claro, em um contexto de aprendizado de máquina, conhecer previamente os padrões significa incorrer em *data snooping*, ou "bisbilhotagem" dos dados. Isso pode ser feito, porém é preciso máximo cuidado ao garantir que os padrões inferidos pelo ser humano tenham real significância estatística. Do contrário, está-se apenas induzindo o algoritmo a aprender um padrão percebido em uma amostra que não reflete o seu conjunto universo ou espaço amostral, pouco ou nada agregando à capacidade de generalização do modelo.

Feitas essas observações, podemos definir o *fluxo* como um conjunto de viagens caracterizado por:

- uma *região de origem* delimitada

- uma *região de destino* delimitada
- um *período de tempo* delimitado

A contagem das viagens que coincidem com um fluxo é a variável alvo do modelo: as viagens existentes são contadas para treinamento do algoritmo de aprendizado de máquina e preditas para novas regiões, em dados filtros de tempo.

3.1.2 Método de amostragem

Os modelos deste trabalho em específico foram baseados em uma divisão da região geográfica em uma grade uniforme (Figura 3.1). Cada região ou *célula* da grade é um possível concentrador de origens ou destinos de viagens. No caso dos sistemas Bluebikes e Indego, podem ser descartadas as células que não possuem estações do sistema em sua área.

A granularidade e o posicionamento da grade podem impactar os fluxos que são determinados. Quanto ao primeiro aspecto, foi escolhido um tamanho de célula com lado de aproximadamente 650 metros, porém diferentes modelos para diferentes granularidades poderiam ser testados. Quanto à posição, ela tem um aspecto arbitrário no sentido de as células poderem tanto conter quanto cortar áreas importantes, de alta concentração de viagens. Esse efeito é de alguma forma aleatório, pois não se pode esperar que uma grade uniforme corresponda exatamente às regiões chave nos aspectos de mobilidade, socioeconômicos e outros. Para mitigar o problema, foram calculadas amostras de fluxos utilizando 2 diferentes posicionamentos da grade, de forma que possa ser explorada maior variedade de distribuição das variáveis no espaço geográfico.

Também, a separação temporal foi feita nas dimensões:

- Mês do ano (agregador)
- Tipos de dias
  - Dias de trabalho
  - Fins de semana e feriados
- Períodos do dia
  - Manhã: entre 7:00 e 9:00
  - Horário de almoço: entre 11:00 e 13:00
  - Final do dia: entre 17:00 e 19:00

A contagem das viagens foi feita para cada possível tupla *(célula de origem, célula de destino, mês do ano, tipo de dia, período do dia)*. A opção de agregar por mês foi feita para refletir o quão constante é um fluxo ao longo do tempo.

Outro aspecto considerado ao compor a amostragem, foi o viés que a disponibilidade do serviço em determinadas áreas introduz na existência de fluxos de viagens entre duas regiões. A Figura 3.1 mostra quais células da grade traçada em Boston possuem estações e quais não possuem, sendo consideradas, portanto, apenas aquelas que possuem estações.A amostragem de fluxos é realizada somente para as células consideradas, na expectativa de que a variabilidade socioeconômica e da quantidade de viagens nesses espaços seja suficiente para se obter um modelo por aprendizagem de máquina.

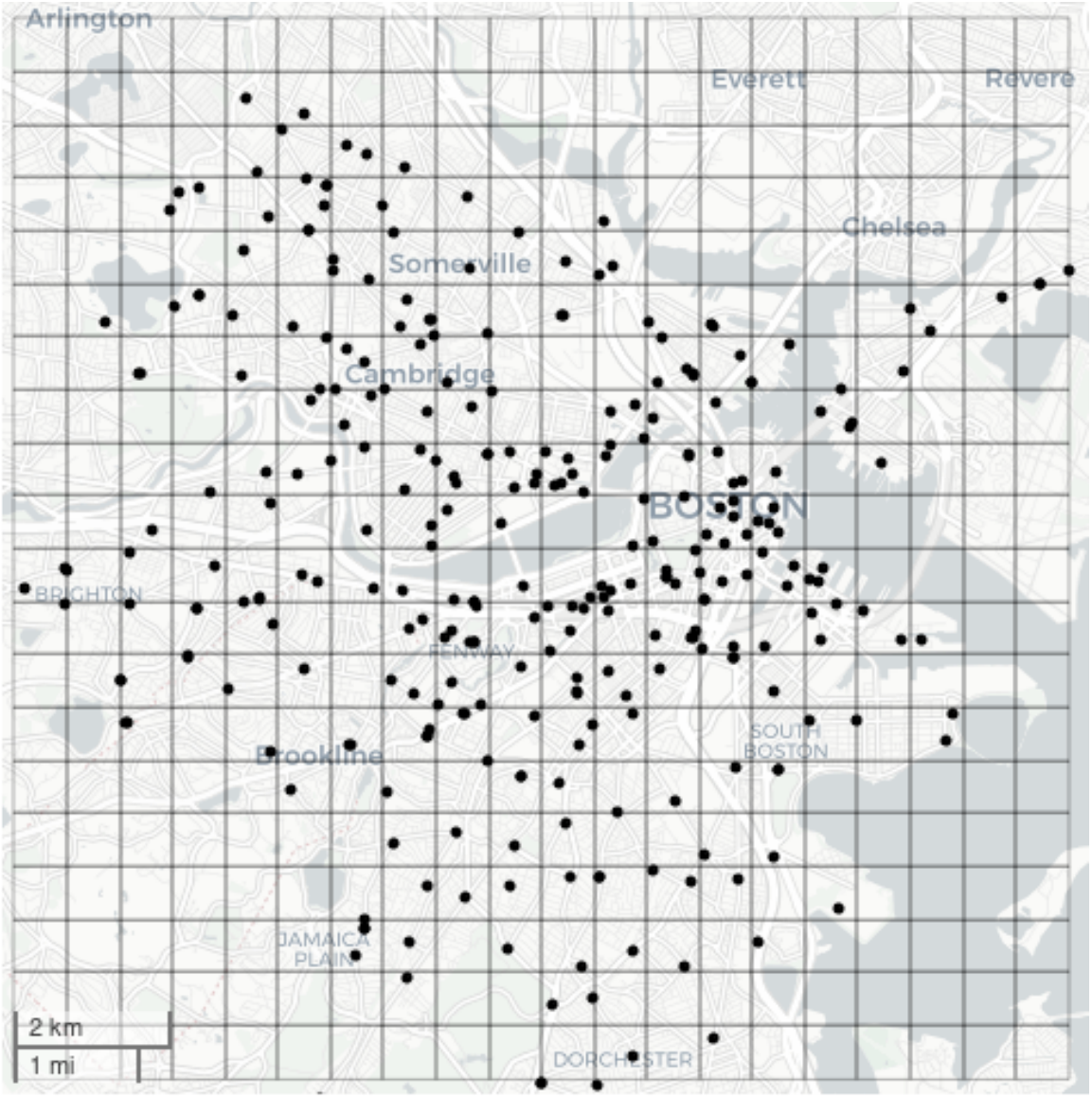

Figura 3.1: *Divisão geográfica em grid da área de atuação do Bluebikes (Boston). Os pontos pretos são as estações (docas) onde bicicletas podem ser retiradas e devolvidas. Fonte: elaborado pelo autor*

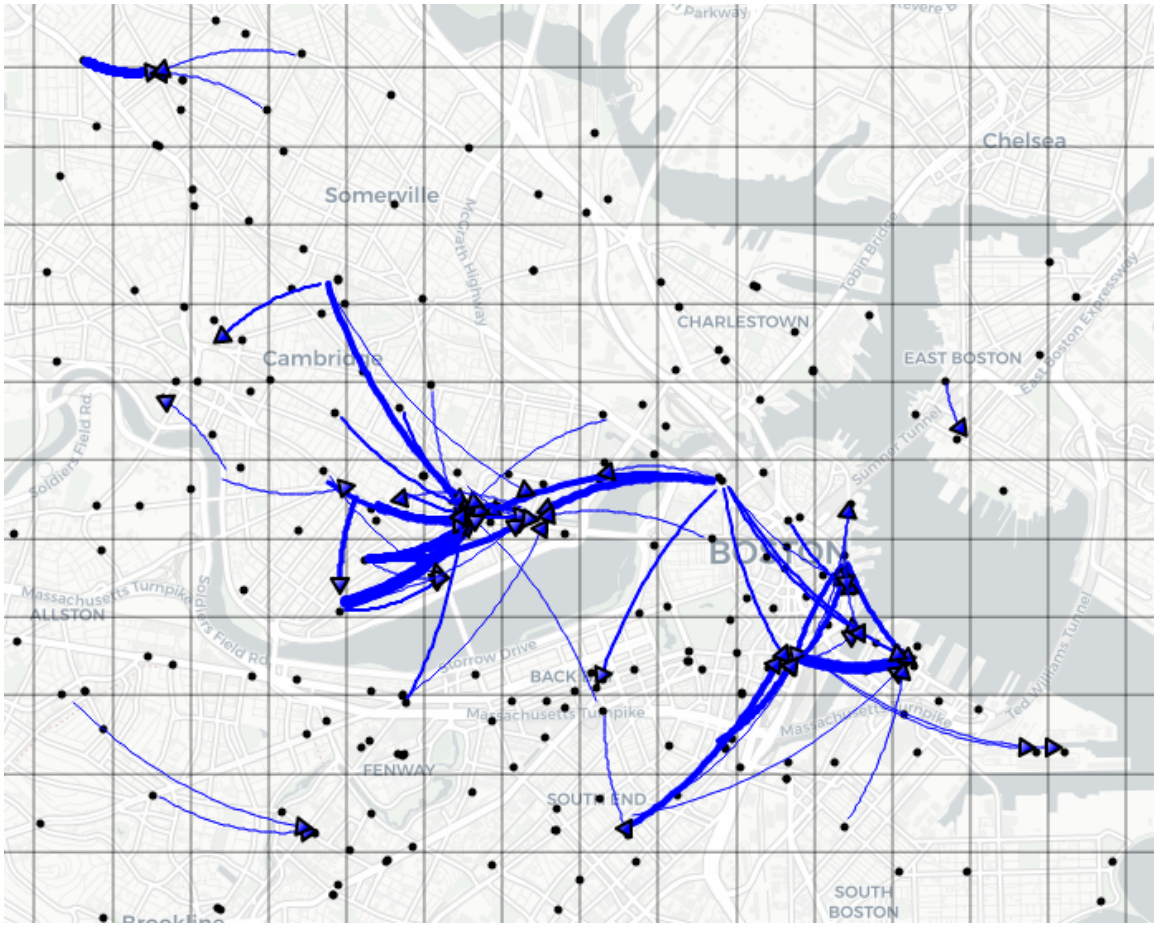

Figura 3.2: *A parcela dos fluxos do Bluebikes (Boston) mais significativos, abrangendo 25% das viagens, para o período da manhã de dias úteis, em março de 2019. Fonte: elaborado pelo autor*

3.1.3 Separação de fluxos por quartis de viagens

A análise dos fluxos de viagens de diferentes Sistemas de Compartilhamento de Bicicletas, realizadas ainda em trabalho preliminar, revela que as viagens são extremamente concentradas geograficamente, isto é, grande parte delas tende a ocorrer entre origens e destinos bem específicos.

Dessa forma, os fluxos encontrados podem ser ordenados do mais significativo para o menos significativo, em número de viagens. Como exemplo, analisando o mês de março de 2019, para os dias de trabalho e o período da manhã (ou seja, fixando a dimensão temporal e variando a dimensão espacial), do sistema Bluebikes, aproximadamente 2,4% dos pares origem-destino são responsáveis por 25% das viagens, e 6,9% por outros 25%. A vasta maioria dos pares origem-destino respondem por contagens de até 17 viagens no mês todo, nos períodos considerados. A tabela 3.1 mostra essa separação das viagens dessa amostragem em quartis.

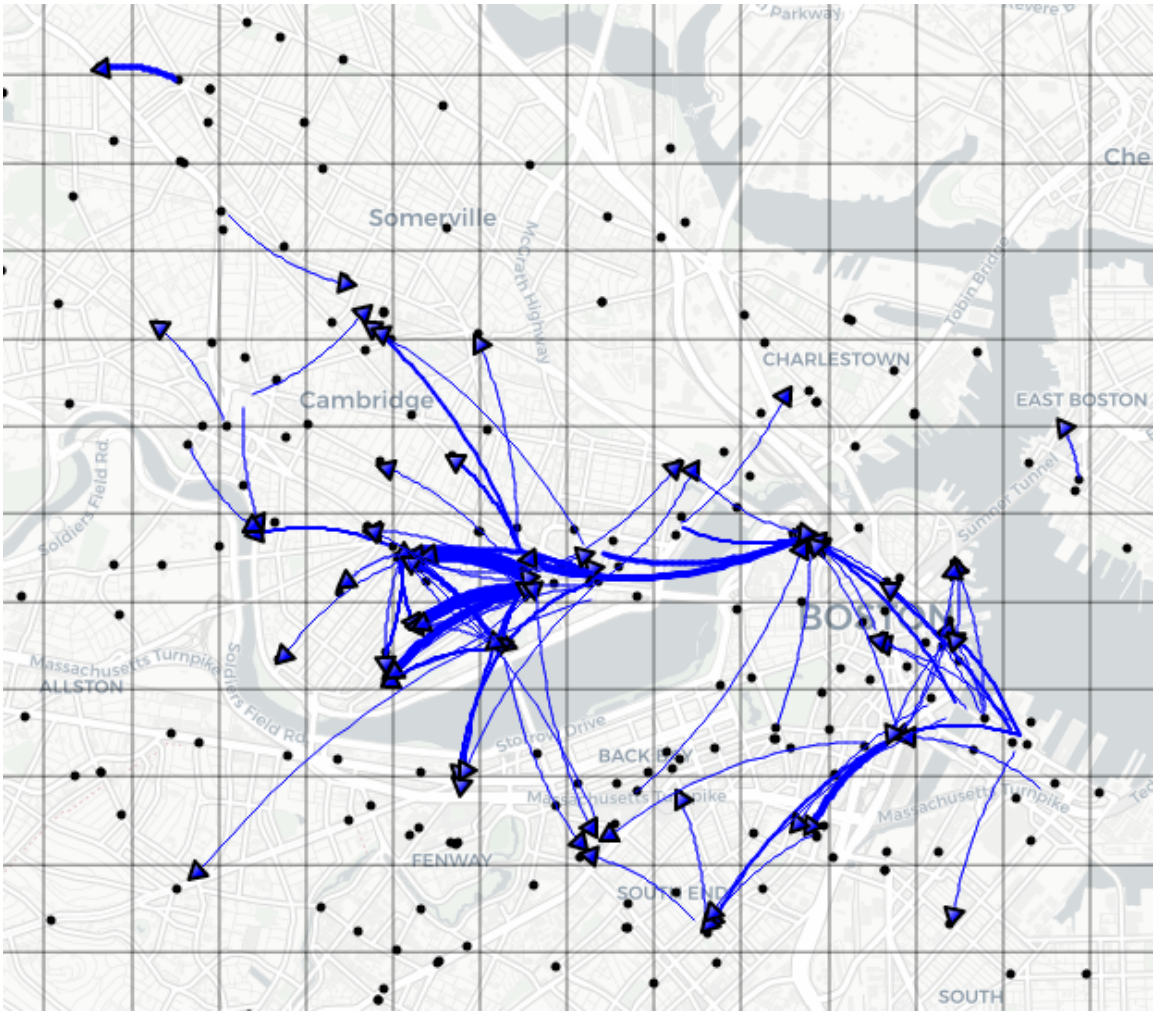

Figura 3.3: *A parcela dos fluxos do Bluebikes (Boston) mais significativos, abrangendo 25% das viagens, para o período do fim de tarde de dias úteis, em março de 2019. Fonte: elaborado pelo autor*

Esses dados mostram que os interessados em uma modelagem de fluxos de origem-destino procurarão prever os poucos fluxos mais significativos, que são uma pequena minoria no conjunto de todos os fluxos mas representam os deslocamentos mais importantes feitos pelos usuários de um Sistema de Compartilhamento de Bicicletas. Os mapas das figuras 3.2 e 3.3 ilustram exemplos de fluxos mais significativos, os menores conjuntos que abrangem 25% de viagens nos períodos considerados. Cada fluxo é representado por uma seta ligando as células de origem e destino. A espessura da seta indica a quantidade de viagens realizada naquele fluxo no período.

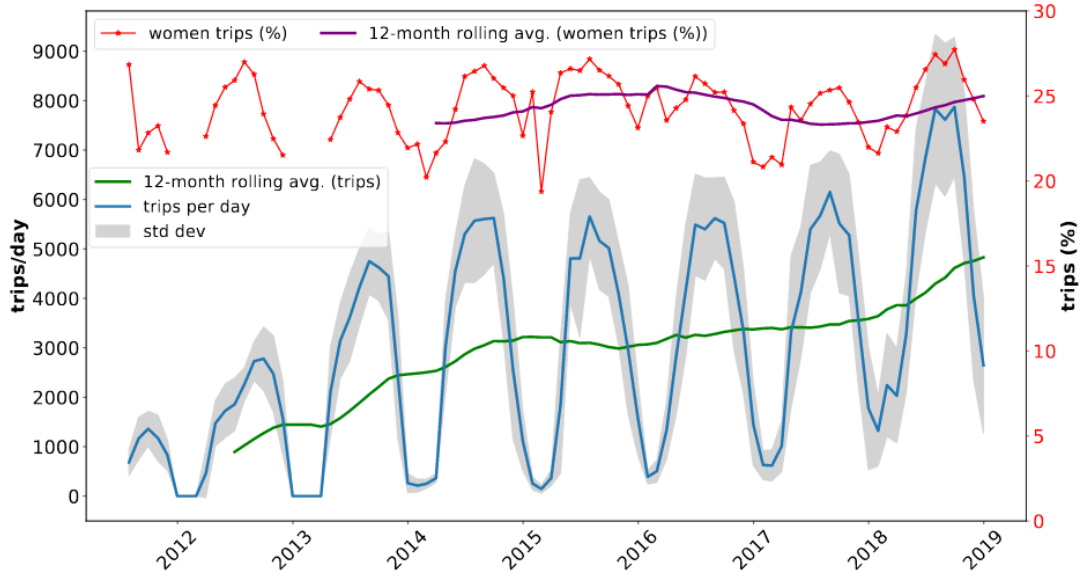

Figura 3.4: *Evolução das viagens do BlueBikes (Boston). Fonte: projeto BikeScience*

Quartil	Mín. viagens	Máx. viagens	Núm. fluxos	% Fluxos
4	44	193	60	2,4
3	17	44	176	6,9
2	8	17	397	15,7
1	1	8	1902	75

Tabela 3.1: *Separação dos fluxos em quartis de viagens do Bluebikes (Boston) para os dias de trabalho e o período da manhã, em março de 2019*

3.2 Fontes de dados

Tendo definido na seção anterior o *fluxo* como o objeto de modelagem da aplicação de aprendizado de máquina sendo desenvolvida, passa-se agora à sua caracterização. Sendo um fluxo um conjunto de viagens realizadas entre duas regiões A e B, em um determinado período do tempo, pode-se dizer que ele possui uma dimensão espacial e uma dimensão temporal. Assim, dados podem ser coletados de forma a caracterizar as regiões de origem e destino e os períodos de tempo e assim gerar um conjunto de atributos ou *features* que alimentarão um algoritmo de aprendizado de máquina.

É preciso ressaltar que este é um aspecto da aplicação que oferece inúmeras oportunidades para melhora da precisão de um modelo gerado por aprendizado de máquina. Aqui entra o conhecimento do negócio e o componente criativo como fundamentais para selecionar indicadores que possam influenciar o uso de um Sistema de Compartilhamento de Bicicletas, bem como fazer perguntas a respeito desses indicadores – uma pergunta interessante feita é: *qual a influência dos indicadores socioeconômicos das diversas regiões,*

obtidos dos dados censitários, no uso de bicicletas pela população?

Esta seção discute de forma isolada cada fonte de dados atualmente usada, e a próxima discute de que formas elas foram integradas e seus dados, relacionados para compôr um único conjunto de dados de caracterização dos fluxos de viagens.

3.2.1 Bluebikes: Sistema de Compartilhamento de Bicicletas de Boston, EUA

O Bluebikes, anteriormente Hubway, é um Sistema de Compartilhamento de Bicicletas público que atua nas cidades de Boston, Brookline, Cambridge, Everett e Somerville, localizadas na região metropolitana de Boston. Possui mais de 300 estações (Figura 3.1), sendo que uma viagem pode começar e terminar em qualquer uma delas, não sendo necessário devolver a bicicleta à mesma estação onde foi retirada. Os usuários podem cadastrar-se online no sistema e adquirir planos de uso, ou pagar por viagem. Também é possível usar o sistema sem cadastro, adquirindo tickets em quiosques localizados junto às estações.

Como sistema público, o Bluebikes disponibiliza abertamente seus dados, os quais podem ser obtidos através do link *System Data* localizado no rodapé de seu site. O histórico de viagens é fornecido em formato tabular, em arquivos do tipo CSV (*comma-separated values*, valores separados por vírgulas), um para cada mês. Os dados já passaram por uma filtragem prévia, não contendo viagens com menos de 1 minuto de duração. Para este trabalho, foram usados os arquivos referentes a abril de 2018 até março de 2019, totalizando 1 853 732 viagens.

A disponibilidade dos dados, à época da coleta, era a partir de julho de 2011. Porém, como a expansão das estações pela área da cidade foi gradual no período, usar um longo histórico de registros poderia enviesar o modelo para favorecer as regiões onde o serviço é oferecido há mais tempo e que, portanto, contam com mais viagens realizadas.

As colunas ou atributos presentes nos arquivos são:

- **tripduration:** duração da viagem (em segundos)
- **starttime:** instante da partida em horário local, formatado como *AAAA-MM-DD HH:MM:SS*
- **stoptime:** instante da chegada em horário local, formatado como *AAAA-MM-DD HH:MM:SS*
- **start station id:** identificador numérico da estação de partida

- **start station name:** nome da estação de partida
- **start station latitude:** coordenada geográfica (latitude) da estação de partida
- **start station longitude:** coordenada geográfica (longitude) da estação de partida
- **end station id:** identificador numérico da estação de chegada
- **end station name:** nome da estação de chegada
- **end station latitude:** coordenada geográfica (latitude) da estação de chegada
- **end station longitude:** coordenada geográfica (longitude) da estação de chegada
- **bikeid:** identificador alfanumérico da bicicleta
- **usertype:** tipo de usuário:
  - *Subscriber:* usuário assinante
  - *Customer:* usuário não assinante
- **birth year:** ano de nascimento do usuário
- **gender:** gêmero do usuário:
  - *0:* desconhecido ou não declarado
  - *1:* masculino
  - *2:* feminino

No período de abril de 2018 a março de 2019, tem-se uma predominância de viagens realizadas por homens (65%), contra 22,4% de viagens de mulheres. Em 12,6% das viagens, o gênero não está declarado. Também, 88,11% são viagens de usuários assinantes (*subscribers*), contra 11,89% de viagens de usuários casuais (*customers*). A Figura 3.5 apresenta as distribuições das idades dos usuários, das durações das viagens, da distâncias percorridas estimadas e das velocidades médias. Como a base de dados não contém o trajeto percorrido, apenas os pontos iniciais e finais, fez-se uso da API GraphHopper para estimar as distâncias entre as estações. Trata-se de um serviço de código aberto capaz de calcular rotas otimizadas para diversos meios de transporte (incluindo bicicletas) a partir de dados do OpenStreetMap.

Essas variáveis, no entanto, não servem aqui como atributos para o treinamento de um algoritmo de aprendizado de máquina, pois não podem ser extrapoladas para novas localidades onde um serviço desse tipo é inexistente. É preciso ater-se às dimensões temporal e espacial, procurando por padrões que variam com essas dimensões. Por exemplo,

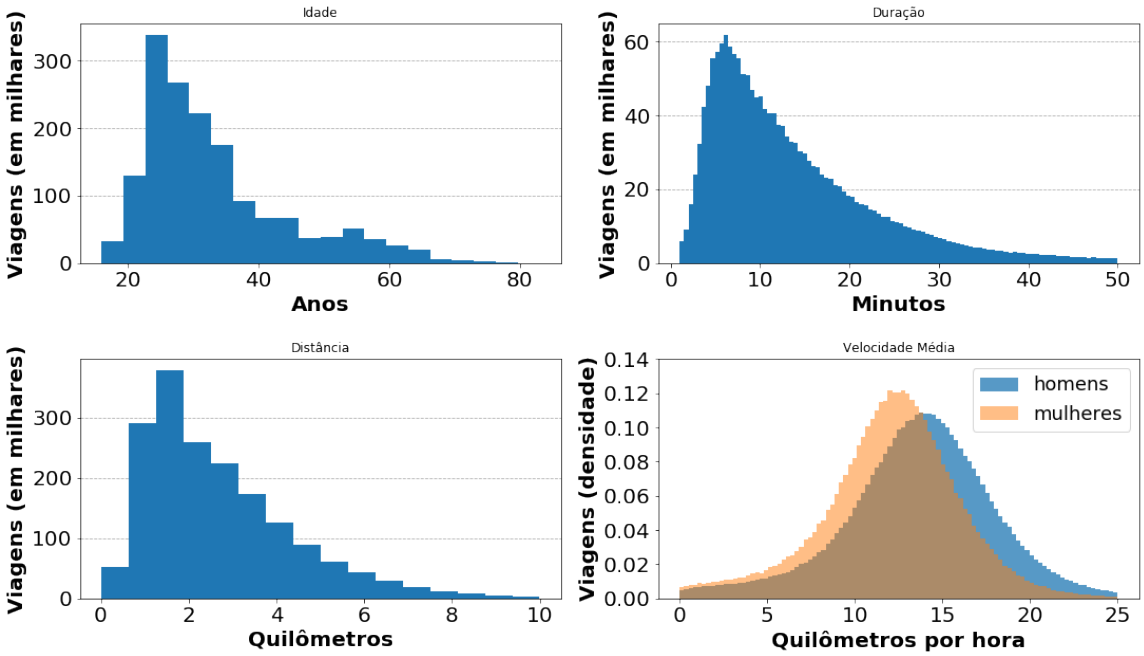

Figura 3.5: *Estatísticas do Bluebikes para o período de abril de 2018 a março de 2019. Fonte: elaborado pelo autor*

comparando as figuras 3.2 e 3.3 é possível ver que muitos fluxos que aparecem no período da manhã com um sentido, aparecem no final da tarde com o sentido contrário. Também, a evolução do número médio de viagens por mês mostra uma variação sazonal significativa (Figura 3.4).

Apesar das estações serem abundantes e estarem bem distribuídas pela área de prestação do serviço, nota-se que o uso delas é bastante desigual, com algumas poucas se destacando. As tabelas 3.4 e 3.3 classificam as estações por uso como origem e destino, respectivamente, mostrando as 20 estações mais frequentes. As tabelas 3.5 e 3.6 apresentam as estatísticas descritivas desses números, respectivamente, para estações de origem e de destino. Sendo visível a influência espacial no uso do sistema, devem-se usar outros conjuntos de dados para caracterizar as regiões onde essas estações se encontram.

Tomando cada estação, podemos verificar em que horários ela é mais usada como origem ou destino de viagens. As figuras 3.6 e 3.7 mostram a média do número de viagens por dia para o período de abril de 2018 a março de 2019, que partem ou chegam, respectivamente, à estação localizada no *Stata Center*, do *Massachussets Institute of Tecnology* (MIT), em cada horário do dia. A análise para diversas estações revela que os horários de pico costumam ser os mesmos e o padrão de uso nos fins de semana é diferente daquele nos dias de trabalho. Assim, considera-se que a agregação de viagens em fluxos separando diferentes períodos do dia e também os dias de trabalho dos fins de semana, permitirá ao algoritmo construir um modelo com maior acurácia.

Estação	Núm. viagens
MIT at Mass Ave / Amherst St	55536
MIT Stata Center at Vassar St / Main St	43815
Central Square at Mass Ave / Essex St	39523
South Station - 700 Atlantic Ave	38278
MIT Pacific St at Purrington St	30955
Kendall T	29737
Nashua Street at Red Auerbach Way	29586
Harvard Square at Mass Ave/ Dunster	27191
Copley Square - Dartmouth St at Boylston St	24881
MIT Vassar St	24671
One Kendall Square at Hampshire St / Portland St	24043
Beacon St at Massachusetts Ave	23366
Ames St at Main St	23070
Boston City Hall - 28 State St	22700
Back Bay T Stop - Dartmouth St at Stuart St	22233
University Park	19109
Christian Science Plaza - Massachusetts Ave at Westland Ave	19009
Central Sq Post Office / Cambridge City Hall at Mass Ave / Pleasant St	18395
Kenmore Square	18279
Cambridge St at Joy St	18085

Tabela 3.2: *Estações mais usadas como origem de viagens no período de abril de 2018 a março de 2019*

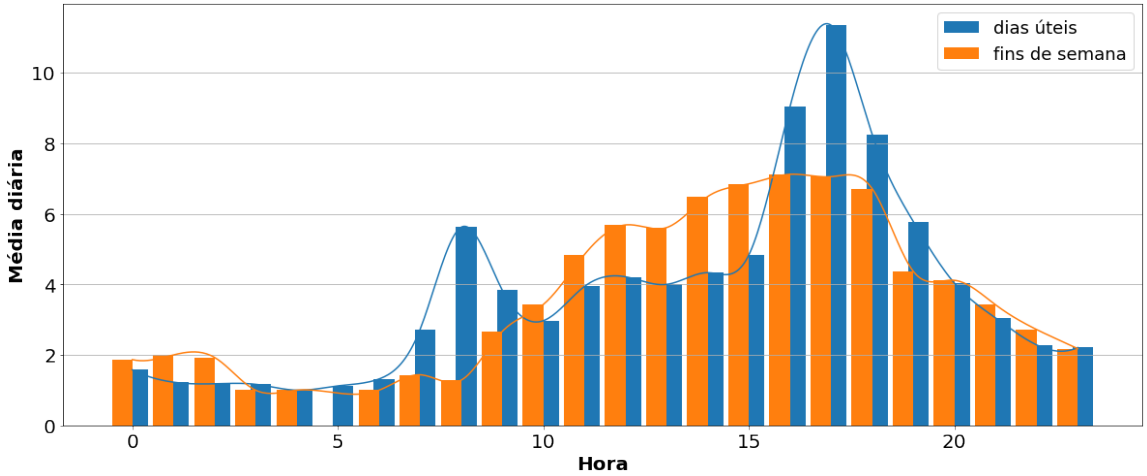

Figura 3.6: *Uso de uma estação localizada no MIT como origem de viagens. Fonte: elaborado pelo autor*

Estação	Núm. viagens
MIT at Mass Ave / Amherst St	50836
MIT Stata Center at Vassar St / Main St	48835
Nashua Street at Red Auerbach Way	39994
Central Square at Mass Ave / Essex St	39540
South Station - 700 Atlantic Ave	36178
MIT Pacific St at Purrington St	28970
Copley Square - Dartmouth St at Boylston St	28238
Harvard Square at Mass Ave/ Dunster	27829
Kendall T	27678
Ames St at Main St	23528
MIT Vassar St	23322
Boston City Hall - 28 State St	21967
One Kendall Square at Hampshire St / Portland St	21777
Back Bay T Stop - Dartmouth St at Stuart St	21150
Beacon St at Massachusetts Ave	21096
Christian Science Plaza - Massachusetts Ave at Westland Ave	19546
Kenmore Square	18724
Cambridge St at Joy St	18118
Central Sq Post Office / Cambridge City Hall at Mass Ave / Pleasant St	17907
Charles Circle - Charles St at Cambridge St	17826

Tabela 3.3: *Estações mais usadas como destino de viagens no período de abril de 2018 a março de 2019*

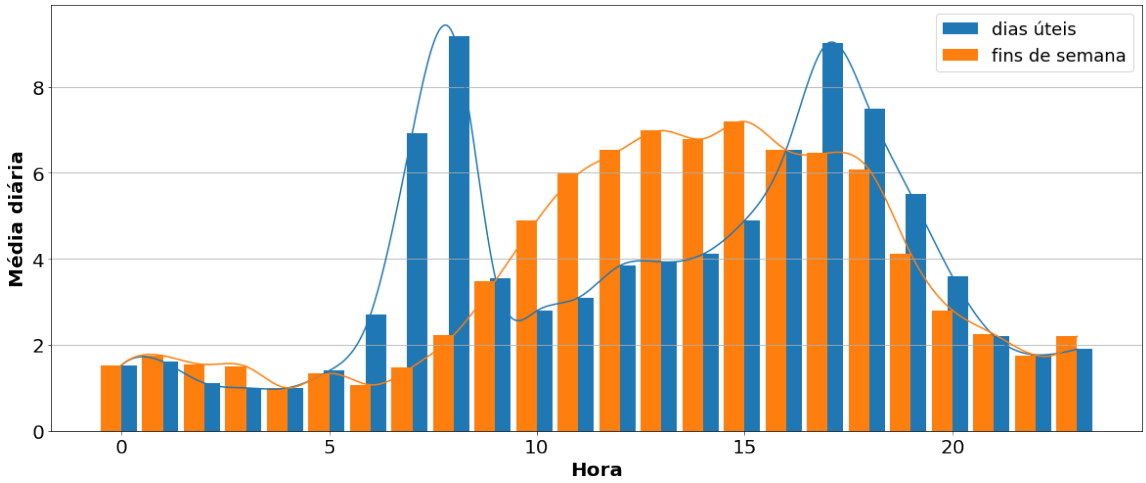

Figura 3.7: *Uso de uma estação localizada no MIT como destino de viagens. Fonte: elaborado pelo autor*

Estação	Núm. viagens
MIT at Mass Ave / Amherst St	55536
MIT Stata Center at Vassar St / Main St	43815
Central Square at Mass Ave / Essex St	39523
South Station - 700 Atlantic Ave	38278
MIT Pacific St at Purrington St	30955
Kendall T	29737
Nashua Street at Red Auerbach Way	29586
Harvard Square at Mass Ave/ Dunster	27191
Copley Square - Dartmouth St at Boylston St	24881
MIT Vassar St	24671
One Kendall Square at Hampshire St / Portland St	24043
Beacon St at Massachusetts Ave	23366
Ames St at Main St	23070
Boston City Hall - 28 State St	22700
Back Bay T Stop - Dartmouth St at Stuart St	22233
University Park	19109
Christian Science Plaza - Massachusetts Ave at Westland Ave	19009
Central Sq Post Office / Cambridge City Hall at Mass Ave / Pleasant St	18395
Kenmore Square	18279
Cambridge St at Joy St	18085

Tabela 3.4: *Estações mais usadas como origem de viagens no período de abril de 2018 a março de 2019*

Estatística	Valor
Núm. estações	336
Média	5517
Desvio padrão	7498
Mínimo	1
1º quartil (25%)	496,2
2º quartil (50%)	2399
3º quartil (75%)	8348,2
Máximo	55536

Tabela 3.5: *Estatísticas descritivas para o número de viagens que partem de cada estação, no período de abril de 2018 a maço de 2019*

Estatística	Valor
Núm. estações	336
Média	5517
Desvio padrão	7549,3
Mínimo	1
1° quartil (25%)	479,8
2° quartil (50%)	2440
3° quartil (75%)	8712,8
Máximo	50836

Tabela 3.6: *Estatísticas descritivas para o número de viagens que chegam em cada estação, no período de abril de 2018 a maço de 2019*

3.2.2 Indego: Sistema de Compartilhamento de Bicicletas da Filadélfia, Estados Unidos

Com o fim de garantir uma maior capacidade de generalização do modelo e evitar que fique enviesado para as características de Boston, o sistema Indego, da cidade da Filadélfia, foi escolhido para complementar os dados do Bluebikes. Assim como este, seus dados são abertos.

Com mais de uma cidade na modelagem, é possível:

- Obter um modelo a partir de uma cidade e testar com outra. Isto é útil para análise de diferenças e semelhanças entre as características de cada local e, assim, ter ideias de modelagem que capturem essas diferenças.
- Obter um modelo conjunto, supostamente mais robusto, a partir de dados de cidades com diferentes características, possivelmente enriquecendo-o com features relativas à cidade.

O conjunto de dados de viagens do Indego apresenta os seguintes atributos:

- **trip_id:** identificador da viagem
- **duration:** duração da viagem em minutos
- **start_time:** instante de início da viagem em horário local
- **end_time:** instante de término da viagem em horário local
- **start_station:** identificador da estação de partida
- **start_lat:** latitude da estação de partida
- **start_lon:** longitude da estação de partida

- **end_station:** identificador da estação de chegada
- **end_lat:** latitude da estação de chegada
- **end_lon:** longitude da estação de chegada
- **bike_id:** identificador da bicicleta
- **plan_duration:** duração do plano adquirido em dias; zero em caso de viagem única
- **trip_route_category:** "Round Trip" se a viagem iniciou e terminou na mesma estação; "One Way" caso contrário
- **passholder_type:** nome do plano adquirido pelo usuário
- **bike_type:** tipo da bicicleta usada (normal ou elétrica)

Da mesma forma que com os dados do Bluebikes, para a modelagem de fluxos é desejável ater-se às dimensões temporal e espacial. Assim, variáveis que idenificam bicicletas, tipos de plano ou usuário podem ser descartadas.

3.2.3 US Census: dados socioeconômicos

Como primeira forma de caracterizar o espaço geográfico, foram coletados dados socioeconômicos levantados pelo *United States Census Bureau* (https://www.census.gov/) referentes a população, sexo, idade, escolaridade e renda. Dentre os programas de pesquisa existentes, em particular, o *American Community Survey (ACS)* fornece estimativas anuais dos indicadores populacionais, oferecendo uma visão das mudanças em andamento antes que os resultados dos censos deceniais sejam publicados – o próximo censo será publicado em 2020.

A API disponível (Bureau, 2019) documenta fartamente os conjuntos de dados disponíveis (https://api.census.gov/data.html). Na relação de *datasets*, é importate atentar-se para a granularidade da divisão do espaço geográfico para a qual um *dataset* está disponível, o que pode ser conferido no link *geographies* ao lado de cada dataset.

Foi procurado um dataset que apresentasse dados, ou ao menos estimativas, em uma base mais atualizada. Como este trabalho foi realizado próximo à publicação do censo de 2020, os dados do último censo (2010) poderiam não refletir adequadamente as condições socioeconômicas do período em que foram considerados os fluxos de ciclistas, isto é, de abril de 2018 a março de 2019. Também, procurou-se uma divisão geográfica o mais granular possível, de forma a corresponder mais aproximadamente à divisão em grade usada na separação das viagens de bicicleta em fluxos.

Foi escolhido o *dataset American Community Survey 5-Year Data* (Bureau, 2018) para o ano de 2017. Trata-se de uma compilação dos dados do ACS para períodos de 5 anos, sendo o último o ano de referência. Mais importante, o **acs5** disponibiliza dados no nível geográfico do *tract*, algo análogo ao setor censitário do IBGE. O *tract* do censo americano (Figura 3.8) é uma área de tamanho variável, que costuma abranger poucos quarteirões, sendo a divisão mais próxima em tamanho da que foi usada na grade da modelagem. As bases de dados geoespaciais (*shapefiles*) foram obtidas em https://www.census.gov/cgi-bin/geo/shapefiles/index.php.

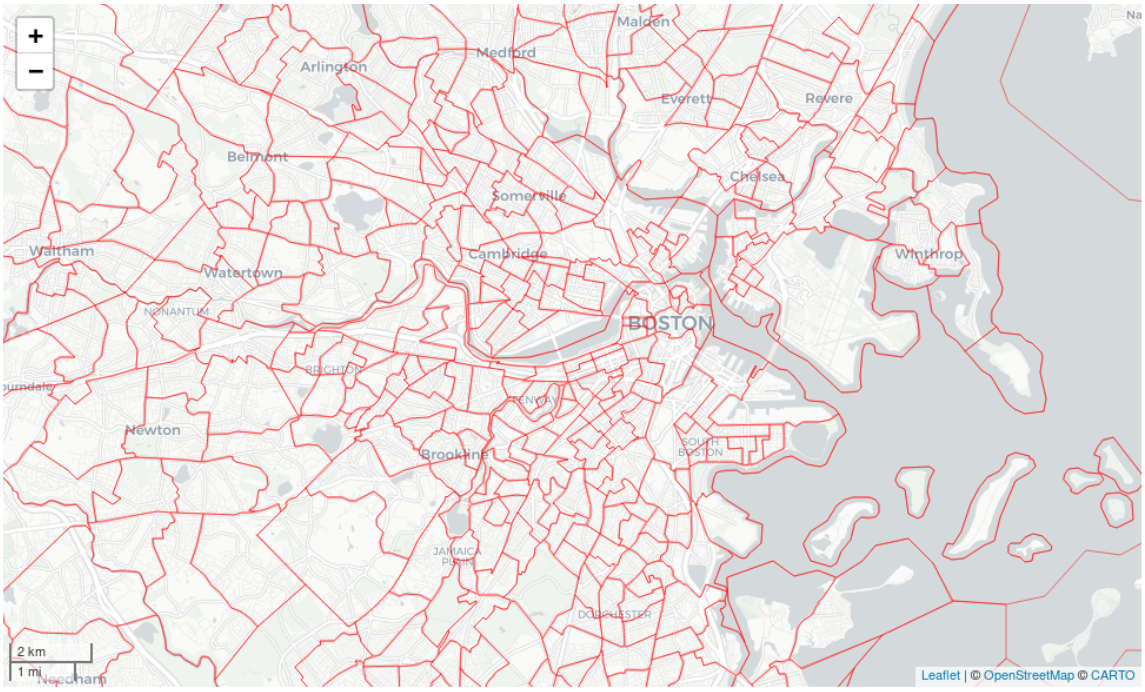

Figura 3.8: *Regiões censitárias (census tracts), para o ano de 2017. Fonte: elaborado pelo autor*

O conjunto de dados **acs5** possui cerca de 64000 variáveis[1]. Esse alto número reflete as diferentes combinações de sexo, faixa etária, raça, ancestralidade, faixa de renda, escolaridade e muitos outros aspectos. Por exemplo, existe uma variável "masculino" com a contagem de homens em um *tract* (ou qualquer nível geográfico disponível escolhido), outra "masculino, entre 15 e 17 anos" decompondo a contagem mais geral, outra "masculino, entre 15 e 17 anos, brancos" e assim por diante. A Tabela 3.7 lista as variáveis escolhidas para a caracterização socioeconômica do espaço.

Na subseção 3.3.2 é discutido como esse conjunto de dados é integrado ao modelo de fluxos de ciclistas.

Variável	Significado

[1] https://api.census.gov/data/2017/acs/acs5/variables.html

B01001_001E	população total
B01001_003E	masculino, abaixo de 5 anos
B01001_004E	masculino, 5 a 9 anos
B01001_005E	masculino, 10 a 14 anos
B01001_006E	masculino, 15 a 17 anos
B01001_007E	masculino, 18 a 19 anos
B01001_008E	masculino, 20 anos
B01001_009E	masculino, 21 anos
B01001_010E	masculino 22 a 24 anos
B01001_011E	masculino, 25 a 29 anos
B01001_012E	masculino, 30 a 34 anos
B01001_013E	masculino, 35 a 39 anos
B01001_014E	masculino, 40 a 44 anos
B01001_015E	masculino, 45 a 49 anos
B01001_016E	masculino, 50 a 54 anos
B01001_017E	masculino, 55 a 59 anos
B01001_018E	masculino, 60 a 61 anos
B01001_019E	masculino, 62 a 64 anos
B01001_020E	masculino, 65 a 66 anos
B01001_021E	masculino, 67 a 69 anos
B01001_022E	masculino, 70 a 74 anos
B01001_023E	masculino, 75 a 79 anos
B01001_024E	masculino, 80 a 84 anos
B01001_025E	masculino, a partir de 85 anos
B01001_027E	feminino, abaixo de 5 anos
B01001_028E	feminino, 5 a 9 anos
B01001_029E	feminino, 10 a 14 anos
B01001_030E	feminino, 15 a 17 anos
B01001_031E	feminino, 18 a 19 anos
B01001_032E	feminino, 20 anos
B01001_033E	feminino, 21 anos
B01001_034E	feminino, 22 a 24 anos
B01001_035E	feminino, 25 a 29 anos
B01001_036E	feminino, 30 a 34 anos
B01001_037E	feminino, 35 a 39 anos
B01001_038E	feminino, 40 a 44 anos
B01001_039E	feminino, 45 a 49 anos

B01001_040E	feminino, 50 a 54 anos
B01001_041E	feminino, 55 a 59 anos
B01001_042E	feminino, 60 a 61 anos
B01001_043E	feminino, 62 a 64 anos
B01001_044E	feminino 65 a 66 anos
B01001_045E	feminino, 67 a 69 anos
B01001_046E	feminino, 70 a 74 anos
B01001_047E	feminino, 75 a 79 anos
B01001_048E	feminino, 80 a 84 anos
B01001_049E	feminino, a partir de 85 anos
B19301_001E	renda *per capita* nos últimos 12 meses
B15002_003E	masculino, sem escolarização
B15002_004E	masculino, com até o 4º ano escolar
B15002_005E	masculino, entre o 5º e o 6º ano escolar
B15002_006E	masculino, entre o 7º e o 8º ano escolar
B15002_007E	masculino, com o 9º ano escolar
B15002_008E	masculino, com o 10º ano escolar
B15002_009E	masculino, com o 11º ano escolar
B15002_010E	masculino, com o 12º ano escolar, sem diploma
B15002_011E	masculino, com ensino médio ou equivalente
B15002_012E	masculino, com menos de 1 ano de faculadade
B15002_013E	masculino, mais de 1 ano de faculdade, sem graduação
B15002_014E	masculino, com grau *associate*[2]
B15002_015E	masculino, com bacharelado
B15002_016E	masculino, com mestrado
B15002_017E	masculino, com graduação em escola profissional
B15002_018E	masculino, com doutorado
B15002_020E	feminino, sem escolarização
B15002_021E	feminino, com até o 4º ano escolar
B15002_022E	feminino, entre o 5º e o 6º ano escolar
B15002_023E	feminino, entre o 7º e o 8º ano escolar
B15002_024E	feminino, com o 9º ano escolar
B15002_025E	feminino, com o 10º ano escolar
B15002_026E	feminino, com o 11º ano escolar
B15002_027E	feminino, com o 12º ano escolar, sem diploma
B15002_028E	feminino, com ensino médio ou equivalente

[2]Graduação para cursos de até 2 anos

B15002_029E	feminino, com menos de 1 ano de faculdade
B15002_030E	feminino, com mais de 1 ano de faculdade, sem graduação
B15002_031E	feminino, com grau *associate*
B15002_032E	feminino, com bacharelado
B15002_033E	feminino, com mestrado
B15002_034E	feminino, com graduação em escola profissional
B15002_035E	feminino, com doutorado

Tabela 3.7: *Variáveis selecionadas do conjunto de dados American Community Survey 5-Year Data de 2017*

3.2.4 Weather API: histórico meteorológico

Dados climáticos foram obtidos da API da *The Weather Company* (Company, 2018) como uma forma de caracterizar a dimensão temporal dos fluxos de viagens de bicicletas. Qualquer que seja o período considerado, instantâneo ou por dia, semana, mês, etc., seus indicadores meteorológicos podem ser agregados e estatísticas podem ser calculadas de forma a caracterizar o período temporal.

A API permite obter tanto dados históricos quanto previsões para até 15 dias a partir do momento da requisição. Os dados históricos são mais ricos em indicadores, fora a disponibilidade de longos períodos de tempo acumulados, o que permite uma análise estatística mais acurada. Isso cria uma questão sobre como extrapolar a modelagem para novos locais: é preciso fazer uma escolha entre usar o histórico (mais rico em dados) ou confiar na significância estatística da previsão do tempo de curto prazo disponível. Neste trabalho, a extrapolação para um novo local é feita considerando os dados históricos do mesmo período da amostragem.

O conjunto de dados apresenta indicadores numéricos e categóricos, bem como descrições textuais e até pictográficas (códigos de ícones indicadores de condição climática). A tabela 3.8 discrimina os indicadores selecionados. Foram escolhidos indicadores que refletem a variação anual, como pode ser percebido no gráfico apresentado na Figura 3.9.

3.2.5 Google Places API: pontos de interesse

Para caracterizar o espaço geográfico e modelar fluxos de ciclistas, é interessante ter em mãos a concentração de *pontos de interesse* nas diferentes regiões de uma cidade. Chamamos aqui de ponto de interesse qualquer local não residencial, para onde as pessoas podem se

Variável	Significado
dewPt	temperaura para formação de orvalho (indicador de umidade)
feels_like	temperatura aparente, ou sensação térmica
heat_index	outra medida de temperatura aparente
precip_hrly	precipitação por hora
rh	umidade relativa
temp	temperatura
vis	visibilidade (afetada pela neblina, poluição ou chuva)
wc	outra medida de temperatura aparente
wspd	velocidade do vento

Tabela 3.8: *Indicadores meteorológicos selecionados do conjunto de dados da The Weather Company*

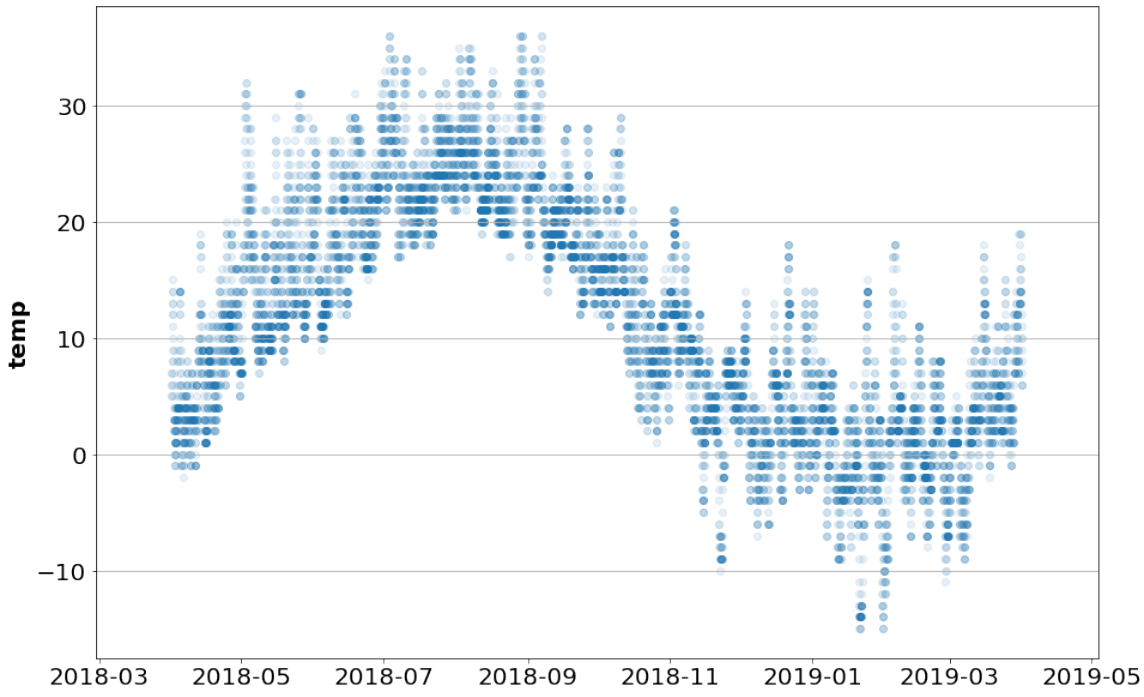

Figura 3.9: *Variação do indicator temp (temperatura) no conjunto de dados do The Weather Company, para Boston, de abril de 2018 a março de 2019. Fonte: elaborado pelo autor*

dirigir para realizar suas atividades, adquirir produtos e serviços, etc. Estabelecimentos comerciais, escolas, igrejas, parques, bares, restaurantes e outros refletem o uso do solo de uma cidade e determinam fortemente os fluxos de mobilidade urbana.

O Google Maps é uma aplicação presente há muito tempo no dia a dia das pessoas como forma de encontrar pontos de interesse. O Google disponibiliza a *Google Places API* (Google, 2019c), a qual permite a consulta aos pontos cadastrados no serviço. Trata-se de uma API paga, parte do serviço *Google Cloud Platform*, o qual oferece um período de avaliação gratuita limitado a 12 meses e a um crédito de US$ 300 (300 dólares americanos) (Google, 2019f).

A licença de uso do serviço permite manter as localizações (latitudes e longitudes) dos pontos em cache por até 30 dias (Google, 2019d). Como o objetivo neste trabalho é realizar uma contagem dos pontos de interesse de cada tipo, por regiões do espaço (as células da grade), isso não é um empecilho, sendo o tempo limite mais que suficiente para a realização desse processamento.

Os pontos cadastrados são classificados em categorias (Google, 2019b), as quais são mostradas na Tabela 3.9. Um ponto pode pertencer a uma ou mais dessas categorias. Essa categorização por si só já oferece um enriquecimento da caracterização do espaço: ao invés de registrar que "há x pontos de interesse em uma célula", pode-se dizer "há i restaurantes, j escolas, k estações de ônibus...".

Para a coleta dos dados, foi escolhida a API JavaScript (Google, 2019a) pelo suporte documentado à delimitação de uma área retangular para a realização de consultas de pontos cadastrados. Foi desenvolvida uma pequena aplicação web composta por uma página HTML, responsável por disparar as consultas via JavaScript, e um *endpoint* para receber e armazenar os dados. Por existir um limite de até 60 pontos devolvidos por consulta, uma grade de alta granularidade foi gerada e fornecida como entrada para a aplicação de forma que, para cada pequena célula, seja disparada uma consulta. Também, como há um limite diário de consultas que podem ser realizadas por dia, a aplicação deve poder retomar a coleta a partir de uma célula arbitrária quando acionada. A Figura 3.10 mostra a grade usada para a coleta e a concentração de pontos coletados em cada célula.

3.2.6 Google Elevations API: relevo e altitude

Assim como a *Places API*, a *Elevations API* é parte do serviço Google Cloud Platform, sendo coberta pelas mesmas condições de serviço descritas para 3.2.5. Trata-se de uma API extremamente simples, a qual recebe a chave de acesso ao serviço e uma lista de coor-

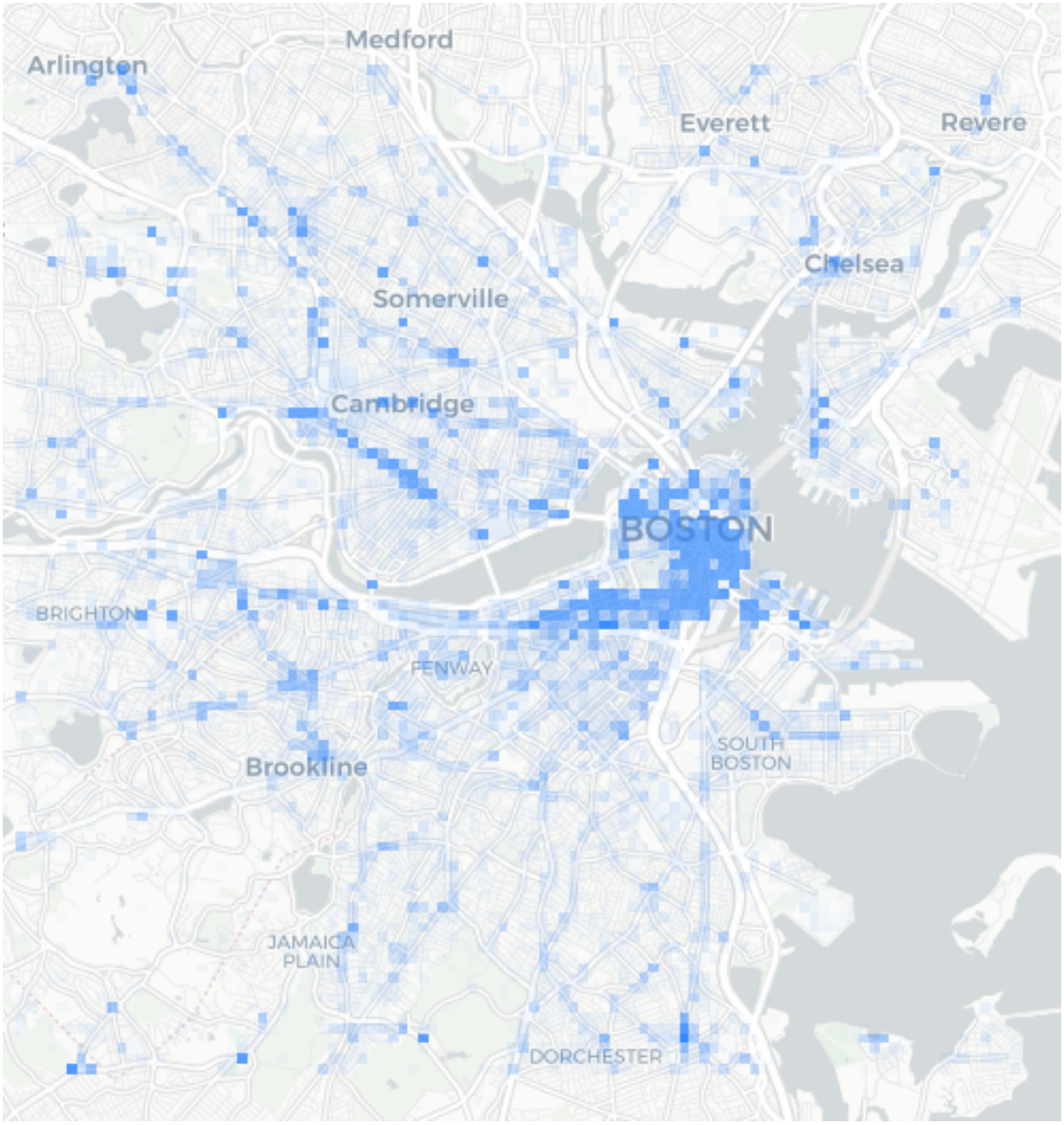

Figura 3.10: *Coleta de pontos de interesse através da Google Places API. As áreas em tons mais escuros de azul são as que devolveram maior quantidade de locais. Fonte: elaborado pelo autor*

accounting	city_hall	laundry	post_office
airport	clothing_store	lawyer	real_estate_agency
amusement_park	convenience_store	library	restaurant
aquarium	courthouse	liquor_store	roofing_contractor
art_gallery	dentist	local_government_office	rv_park
atm	department_store	locksmith	school
bakery	doctor	lodging	shoe_store
bank	electrician	meal_delivery	shopping_mall
bar	electronics_store	meal_takeaway	spa
beauty_salon	embassy	mosque	stadium
bicycle_store	fire_station	movie_rental	storage
book_store	florist	movie_theater	store
bowling_alley	funeral_home	moving_company	subway_station
bus_station	furniture_store	museum	supermarket
cafe	gas_station	night_club	synagogue
campground	gym	painter	taxi_stand
car_dealer	hair_care	park	train_station
car_rental	hardware_store	parking	transit_station
car_repair	hindu_temple	pet_store	travel_agency
car_wash	home_goods_store	pharmacy	veterinary_care
casino	hospital	physiotherapist	zoo
cemetery	insurance_agency	plumber	
church	jewelry_store	police	

Tabela 3.9: *Categorias de pontos cadastrados na Google Places API (Google, 2019b)*

denadas geográficas e devolve as altitudes dos pontos em metros (Google, 2019e).

Uma diferença significativa de altitude pode influir na existência de um fluxo de ciclistas entre duas regiões A e B. Por exemplo, em São Paulo, existem fluxos que partem da Avenida Paulista e vão em direção ao Parque do Ibirapuera, porém fluxos no sentido contrário são praticamente inexistentes porque as pessoas teriam que pedalar ladeira acima. Dessa forma, caracterizamos as regiões de origem e destino pela altitude de algum ponto da célula, por exemplo, seu centroide.

3.2.7 OpenStreetMap e GraphHopper: estrutura cicloviária e rotas

Os dados de viagens de um Sistema de Compartilhamento de Bicicletas possuem um viés intrínseco em favor das áreas onde a infraestrutura para pedalar já é existente. Com o objetivo de capturar esse viés, dados geoespaciais da estrutura cicloviária, isto é, os traçados de vias para bicicletas (segregadas ou não) foram coletados. Foi utilizada a biblioteca *OSMnx*, a qual coleta e processa grafos de vias da API aberta OpenStreetMap.

Tendo em mãos os traçados, fica a pergunta: existe infraestrutura cicloviária no caminho entre duas células da grade? Serviços de traçado de rotas como o Google Maps ou o GraphHopper (https://www.graphhopper.com/) permitem traçar um caminho entre dois pontos, adequado e otimizado para ser feito por transporte público, carro particular, bicicleta ou caminhada a pé. O GraphHopper foi escolhido por também usar os dados do OpenStreetMap, permitindo corresponder geometricamente os traçados das rotas sugeridas, otimizadas para bicicleta, e da infraestrutura cicloviária.

3.3 Integração

Tendo definido quais os conjuntos de dados comporão o modelo em um primeiro momento, surge a questão: como integrá-los? Dados de fontes diferentes dificilmente possuem atributos comuns, ainda mais em se tratando de dados com finalidades a princípio não relacionadas.

Como mencionado em 3.2, os fluxos de viagens possuem uma dimensão temporal, formada pelo mês, período do dia (manhã, hora de almoço, fim de tarde) e tipo de dia (dia de trabalho, fim de semana ou feriado), e uma dimensão espacial, que é a célula da grade que agrega pontos de partida e chegada das viagens. Os outros conjuntos de dados relacionar-se-ão com os fluxos através de uma dessas dimensões. Em se tratando da dimensão espacial, será necessário lançar mão do GeoPandas e seu recurso de junção espacial. Quanto à dimensão temporal, é preciso agregar dados pelos períodos de tempo desejados, e em seguida corresponder os períodos em dois conjuntos de dados através de uma junção por atributos.

3.3.1 Fluxos de viagens: o início

O principal objeto de modelagem, o *fluxo*, não existe como dado fornecido por algum Sistema de Compartilhamento de Bicicletas. O conjunto de fluxos é uma abstração de um conjunto de viagens, sendo determinado conforme descrito em 3.1.

O primeiro passo é a determinação de uma grade (Figura 3.1), ou qualquer outra divisão do espaço que se achar conveniente, em qualquer granularidade que se desejar. A grade é montada gerando-se uma lista de objetos *box* (retângulos) da biblioteca *Shapely* e em seguida convertendo essa lista em um *dataframe* do *GeoPandas* (*geodataframe*). Cada célula possui um identificador formado pelos atributos i e j, identificando as células da mesma forma que em uma matriz matemática.

Em seguida, agregam-se as viagens que se iniciam e terminam em cada célula através

de junção espacial. A operação cruza o conjunto de viagens com o conjunto de células, determinando que pontos se localizam em que células. Extrair antes um conjunto de estações acelera a execução desse passo, pois não é preciso considerar um ponto para cada viagem.

A partir da grade já é possível caracterizar um fluxo, calculando a distância entre duas células. São tomados os centroides das células e a distância entre cada par é calculada pelo método de *haversine*.

Para a execução do aprendizado de máquina, duas grades foram usadas, com um ligeiro deslocamento entre elas, de forma a melhor amostrar a distribuição um tanto quanto arbitrária de elementos do espaço pelas células. Somente fluxos entre células da mesma grade são considerados. Cada grade é identificada por um atributo $placement\_id$.

Por fim, para cada par $(origem, destino, grade)$ de celulas, e para cada período *(mês, período do dia, tipo de dia)* as viagens são contadas. O conjunto de fluxos é um *dataset* como o mostrado na Figura 3.11. O sufixo *_start* identifica as células de origem, e o sufixo *_end* identifica as células de destino.

	i_start	j_start	i_end	j_end	placement_id	trip counts	period	distance	month	weekend_or_holiday
0	0	3	0	5	0	0.0	1	2661.228358	2018-05-01	0
1	0	3	0	6	0	0.0	1	3991.842507	2018-05-01	0
2	0	3	1	3	0	0.0	1	1335.475153	2018-05-01	0
3	0	3	1	4	0	0.0	1	1885.123403	2018-05-01	0
4	0	3	1	5	0	0.0	1	2977.293951	2018-05-01	0

Figura 3.11: *Amostragem de um conjunto de dados de fluxos de viagens e seus atributos. Fonte: elaborado pelo autor*

3.3.2 Integrando dados censitários

Como descrito em 3.2.3, os indicadores socioeconômicos do censo são dados por setores ou *tracts*, regiões de tamanho variável (Figura 3.8). Isso cria um problema: como representar esses indicadores em uma grade uniforme, de células retangulares?

Certamente os fluxos de viagens poderiam ser calculados sobre os setores censitários, isto é, usando-os como regiões de origem e destino de viagens. No entanto, a grade é algo que permite modelar granularidades diferentes, ou seja, computar fluxos entre regiões menores ou maiores conforme a distância mínima que se deseja considerar um fluxo.

Também, as células intersectar-se-ão com os setores do censo de maneira arbitrária: uma célula pode conter setores menores, um setor maior pode conter células, um setor pode

ter partes de células e vice-versa. Pensando nisso, foi elaborada uma forma de *distribuir* os setores pelas células proporcionalmente às suas áreas de interseção. Usando diferentes posicionamentos da grade, podem-se amostrar diferentes distribuições dos setores pelas células de cada grade.

Para cada setor, determina-se o percentual de sua área que está contido em cada célula através de processamento geoespacial. Para cada indicador socioeconômico i (Tabela 3.7) de um setor S, se uma célula C possui $p\%$ da área de S, estima-se que ela possui $p\%$ do valor de i. Todos os setores presentes em uma célula (toda a área ou parte), com seus valores proporcionais, são agregados, caracterizando a célula com o mínimo, o máximo, a média e o desvio padrão dos valores i de cada setor. Assim, cada indicador socioeconômico gera 8 atributos para um fluxo: as quatro estatísticas agregadas, para ambas as células de origem e destino.

3.3.3 Integrando pontos de interesse

À primeira vista parece simples integrar o conjunto de pontos de interesse ao conjunto de fluxos pela dimensão espacial. Bastaria contar os pontos de cada tipo por célula da grade, e caracterizar as células de origem e destino do fluxo com as contagens.

No entanto, isso leva a imprecisões devido à arbitrariedade intrínseca da grade. Uma célula pode cortar uma área de alta concentração de pontos de interesse, como um bairro comercial. Uma pessoa pode retirar e deixar a bicicleta em uma estação, sendo que sua origem ou destino é um local a alguns metros dali. Viagens que começam e terminam em regiões nas bordas das células podem relacionar-se a pontos de interesse nas células vizinhas.

Da mesma forma que com os setores do censo (ver tópico anterior), é preciso encontrar uma forma de *distribuir* os pontos de interesse pelas células. Aqui também, uma variedade de posicionamento da grade pode fornecer maior amostragem de distribuições dos pontos pelas células.

Primeiro é preciso considerar um pouco além do exterior da célula. A partir do centroide do retângulo, um círculo de raio 600m é obtido por uma operação denominada *buffering* geodésico[3] (Figura 3.12) (Flater, 2011). Esse círculo será denominado área de alcance da célula. Em seguida, realiza-se a junção espacial do conjunto de círculos com o conjunto de

[3] O GeoPandas oferece uma operação de *buffering* mais simplificada, somente por geometria euclidiana. Devido à distorção característica de uma projeção cartográfica, o *buffer* gerado pelo GeoPandas na latitude de Boston possui formato ovalado. Sobre isto: https://gis.stackexchange.com/questions/289044/creating-buffer-circle-x-kilometers-from-point-using-python.

pontos de interesse, obtendo a relação de pares (p, c) de pontos p e círculos c tais que $p \in c$. Finalmente, os pontos são rateados pelas células em cuja área de alcance se encontram: se um ponto se encontra na área de alcance de apenas 1 célula, a célula então possui 1 ponto; se está na interseção de 2 áreas de alcance, cada célula tem 1/2 ponto, e assim por diante.

As contagens dos pontos de interesse proporcionalmente distribuídos são realizadas para cada tipo descrito na Tabela 3.9. Para cada tipo, o fluxo é caracterizado com uma contagem para a célula de origem e uma contagem para a de destino.

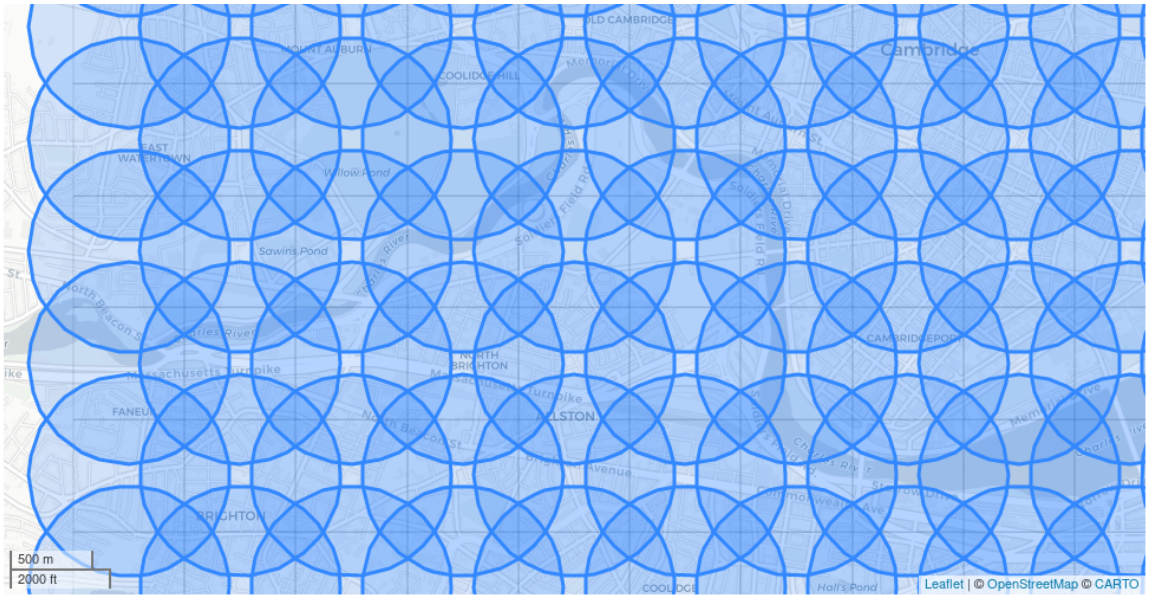

Figura 3.12: *Áreas de alcance das células para o cômputo de pontos de interesse. Fonte: elaborado pelo autor*

3.3.4 Integrando indicadores meteorológicos

O histórico de dados meteorológicos está relacionado com a dimensão temporal dos fluxos de viagens. Um fluxo agrega viagens por períodos de tempo, enquanto o histórico meteorológico apresenta medições periódicas de temperatura, umidade e outros indicadores, descritos na Tabela 3.8. Assim, este também deve ser agregado pelas mesmas variáveis de tempo que os fluxos, a saber: o mês, o período do dia (manhã, hora de almoço, fim do dia) e o tipo do dia (dias de trabalho ou fins de semana e feriados).

Em se tratando de períodos de tempo como um mês, perde-se a precisão das medições instantâneas. O que se pode fazer é agregar as estatísticas descritivas de cada indicador para cada período de tempo. Assim, o fluxo é caracterizado pelo mínimo, máximo, média e desvio padrão de cada indicador meteorológico no período que agrega.

3.3.5 Integrando a estrutura cicloviária

Foi descrito em 3.2.7 a necessidade de se dispor da infraestrutura cicloviária como forma de modelar o viés que ela introduz no conjunto de dados de viagens. Também, foram capturadas rotas entre as células da grade através da API do GraphHopper. As rotas foram obtidas para cada par *(origem, destino)* de células, utilizando seus centroides como pontos de partida e chegada, e otimizadas para bicicleta. Dessa forma, o GraphHopper procura sugerir caminhos sobre a infraestrutura cicloviária, onde disponível. Para saber em que fluxos isso ocorre, faz-se a junção espacial entre o conjunto de vias cicloviárias e o conjunto de rotas.

Ambos os conjuntos de vias para bicicletas e de rotas contém linhas como objetos geoespaciais. Intersectar linhas resulta em objetos ponto, os quais podem ser resultantes de cruzamentos, e não de traçados iguais. Para contornar isso, foi usado o recurso de *buffering* do GeoPandas. O *buffering* cria um polígono ao redor da linha, obtendo uma faixa de alguma largura desejada. Em seguida, faz-se a interseção entre os polígonos e toma-se a área dessa interseção (Figura 3.13).

A disponibilidade de ciclovias para um fluxo é modelada como a razão entre a área total de interseção da sua rota com ciclovias e a área da rota em si, considerando os *buffers* ou faixas poligonais.

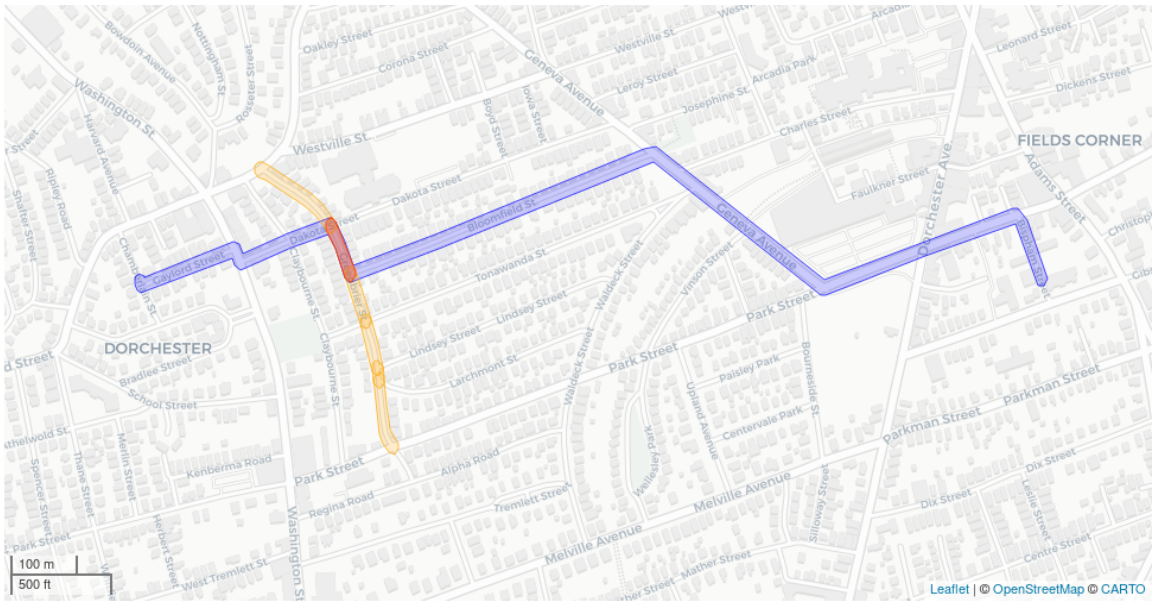

Figura 3.13: *Determinação da disponibilidade de estrutura cicloviária para um fluxo por bufferização (faixa poligonal ao redor do traçado). Em azul, a rota sugerida pelo GraphHopper. Em laranja, um segmento de ciclovia ou ciclofaixa. Em vermelho, a interseção entre ambas. Fonte: Elaborado pelo autor*

3.3.6 Integrando elevações

Tendo sido coletada a elevação ou altitude para cada célula da grade, para caracterizar um fluxo a maneira mais simples seria tomar as altitudes das células de origem e destino. No entanto, como a elevação pertence a um ponto (no caso, o centroide da célula), viagens que começam e terminam próximo às bordas podem ser mais impactadas pela altitude de uma célula vizinha, em especial se o relevo do terreno for bastante acidentado.

Como forma de modelar uma possível variação de altitude interna à célula, para cada uma, é tomada a altitude de suas vizinhas ao norte, ao sul, ao leste e ao oeste. Calcula-se uma média ponderada, de forma que a altitude da célula em si possua peso $0,5$ e a de suas vizinhas, peso $0,125$ ($= 0,5/4$). Ambos os valores, altitude do centroide e média ponderada da célula com suas vizinhas, são usadas para caracterizar a célula. Para o fluxo, os valores são tomados para as células de origem e destino.

3.4 Conjunto de dados de amostra

Após as integrações de diferentes conjuntos de dados brutos, devidamente processados, ao conjunto de fluxos de origem e destino calculados a partir dos dados de viagens do Sistema de Compartilhamento de Bicicletas, obtém-se um conjunto de dados a ser usado como entrada para algum algoritmo de aprendizado de máquina. Esse conjunto de dados é gerado repetidas vezes, conforme ideias de modelagem e implementação vão sendo aplicadas, processo descrito em 3.5.

No momento da entrega do presente trabalho, o conjunto de amostra apresenta o seguinte conjunto de variáveis ou *features* para caracterizar um fluxo:

- Calculadas a partir das viagens: distância entre as células de origem e destino, mês do ano, período do dia, flag indicadora de fim de semana ou feriado **(4 features)**.
- Do censo americano (US Census): 80 indicadores (Tabela 3.7) distribuídos entre os *tracts* que intersectam uma célula. Para cada célula, o máximo, o mínimo, a média e o desvio padrão dos *tracts* (4 estatísticas agregadas). Para cada fluxo, $80 \times 4 \times 2 =$ **640 features**.
- Da Weather API: 9 indicadores (Tabela 3.8) agregados para cada período de tempo no mínimo, máximo, média e desvio padrão (4 estatísticas), resultando **36 features**.
- Da Google Places API: há 90 tipos de pontos de interesse (Tabela 3.9). Cada célula possui uma feature por tipo, totalizando a porção rateada de pontos daquele tipo. Para um fluxo, $90 \times 2 =$ **180 features**.

- Da Google Elevations API: para cada célula, 2 features: a altitude do centroide e a média ponderada das altitudes da célula e seus vizinhos. Para um fluxo, **4 features**.
- Da estrutura cicloviária: **1 feature** representando a porcentagem da área do buffer da rota sugerida pelo GraphHopper para o fluxo que intersecta com buffers da estrutura cicloviária.

Assim, o conjunto de dados de amostragem e caracterização dos fluxos possui 4 + 640 + 36 + 180 + 4 + 1 = 865 features.

3.5 Processo de refinamento do modelo

Obter um modelo preditivo por aprendizado de máquina pode ser uma tarefa extremamente complexa, a depender do problema sendo tratado e, principalmente, do conhecimento que se tem sobre ele. Especialmente em um trabalho de pesquisa como o *BikeScience*, onde se procura avançar a fronteira do conhecimento, hipóteses são formuladas, testadas e ao fim validadas ou descartadas, significando que o software de modelagem deve se adaptar facilmente a novas necessidades.

O trabalho foi realizado em um conjunto de cadernos Python (*Jupyter Notebooks*), estruturados em uma sequência lógica de forma que um caderno realiza seu processamento, exibe informações relevantes na forma de tabelas, gráficos e mapas e grava em disco um conjunto de dados intermediário, que servirá como entrada para algum outro caderno posterior na sequência. Assim, a depender do tipo de refinamento ou modificação que se deseja realizar na modelagem, as partes não afetadas não precisam ser reprocessadas, bastando pegar o resultado pré-gravado.

Os cadernos são numerados de forma que os responsáveis pela coleta de dados sejam identificados como 1.x, os responsáveis por integrar esses dados, como 2.x e os responsáveis por gerar e avaliar modelos como 3.x. Assim, para uma nova coleta de dados é criado um caderno 1.x e sua integração é feita em um caderno 2.x. Mudanças nas integrações exigem somente trabalhar no caderno 2.x. O caderno $2.x - Join - All.ipynb$ é o único que precisa necessariamente ser ajustado e reexecutado a cada mudança em outros, pois é o responsável por gerar o conjunto de dados final para ajuste do modelo.

Os arquivos de saída gerados identificam seus registros de forma a relacioná-los às dimensões temporal e espacial dos fluxos. Conforme visto na Figura 3.11, um fluxo é identificado em sua dimensão espacial pelos atributos *i_start*, *j_start*, *i_end*, *j_end* e *placement_id*, e em sua dimensão temporal pelos atributos *month*, *period* e *weekend_or_holiday*. Um arquivo intermediário pode caracterizar células da grade, apresentando os atributos

i, *j* e e*placement_id* e devendo ser mesclado (Figura 2.7) com o conjunto de fluxos duas vezes, uma para a célula de origem e outra para a célula de destino, ou pode conter toda a identificação espacial ou temporal de um fluxo, bastando mesclá-los pelos atributos correspondentes. Essa tarefa é realizada no caderno $2.x - Join - All.ipynb$.

Os atributos de identificação da dimensão espacial, como se referem a convenções arbitrárias (as grades) e não carregam em si informação relevante para o problema, **são usados somente para identificação e não são considerados como *features* de entrada para um algoritmo de aprendizado de máquina**.

Para cada cidade sendo modelada, um diretório *ml-models* contém os seguintes cadernos:

3.5.1 1.1-Month-Day-Period-Day-Type-Flows.ipynb

Caderno responsável por ler os dados do Sistema de Compartilhamento de Bicicletas e chamar os módulos Python do *BikeScience* para criar as grades e determinar os fluxos. Determina as células a serem consideradas (as que possuem estações) e armazena as possíveis combinações de origem e destino para futuro cálculo de *features* sobre essas células. As viagens são separadas por mês, período do dia e tipo de dia (dias de trabalho e fins de semana ou feriados).

3.5.2 1.2-Points-Of-Interest.ipynb

Como descrito em 3.2.5, uma aplicação web roda uma API JavaScript responsável por coletar pontos de interesse da base de dados do Google, gerando um arquivo de dados brutos. Essa aplicação, localizada no diretório *poi-collection*, possui uma página HTML contendo o JavaScript e um endpoint criado com o framework *Flask* do Python, o qual recebe cada dado coletado e o grava no arquivo.

O caderno processa esse arquivo bruto e gera um arquivo de pontos já coletados com as colunas dos tipos (Tabela 3.9) representando features, pois um ponto pode vir classificado em um ou mais tipos, e facilitando a agregação posterior com as células da grade de fluxos. Também, como a API de coleta do Google impõe uma limitação diária de requisições, o caderno é capaz de atualizar o JavaScript da aplicação web para eliminar as células cujos pontos já foram coletados, impedindo repetições de coleta em caso de erro.

3.5.3 1.3-US-Census.ipynb

Caderno responsável por acessar a Census API e collher os indicadores. Como a API impõe limites na quantidade de variáveis que podem ser coletadas a cada requisição, o caderno trata de coletá-las em blocos e uni-los em seguida. Para trazer o mínimo possível de dados necessários, os recursos de filtragem da API por estado e condado são aplicados. Os indicadores são solicitados no nível de granularidade dos *census tracts* (3.2.3).

Sua saída é um conjunto de dados com as colunas representando os indicadores e as linhas representando os *tracts*, os quais posteriormente são correspondidos com as células por junção espacial.

3.5.4 1.4-Weather-API-Historical-Data.ipynb

Caderno responsável por acessar a Weather API e coletar dados meteorológicos. A API recebe um par $(latitude, longitude)$ (para o qual são fornecidas as coordenadas oficiais da cidade) e o período de tempo, que é seu limitador de requisições – logo, várias coletas são feitas, uma para cada mês. É obtido o conjunto de medições no período, cada uma realizada em diferentes intervalos regulares ao longo de um dia. Portanto, um registro indica que no instante t_1 a variável v_1 vale x_1, outro indica que no instante t_2 o valor de v_2 é x_2 e assim por diante.

As variáveis são inspecionadas, tendo suas medições analisadas ao longo do tempo, e são selecionadas aquelas que apresentam variação sazonal visível e que podem ser obtidas posteriormente com o recurso de previsão do tempo oferecido pela API.

3.5.5 1.6-Bike-Facilities.ipynb

Este caderno realiza acessos à API do GraphHopper, com o objetivo de traçar rotas adequadas para bicicleta entre as células, e cruzá-las geometricamente com a infraestrutura cicloviária. Esta é obtida por meio da biblioteca Python *OSMnx*, a qual acessa o OpenStreetMap, que é a mesma fonte de dados usada pelo GraphHopper.

Tendo o GraphHopper um limite de 500 requisições diárias no plano gratuito, o caderno deve ser capaz de salvar seu progresso e disparar a coleta para os fluxos faltantes. Os fluxos que devem ser calculados são determinados no caderno *1.1-Month-Day-Period-Day-Type-Flows.ipynb* e aqui são cruzados com as rotas já obtidas para determinar o que ainda falta ser coletado.

A qualquer momento (de preferência ao final da coleta), podem ser invocadas as funcionalidades de *buffering* e *junção espacial* por interseção (3.3.5) do GeoPandas para

determinar que rotas intersectam com que fluxos. O resultado é um arquivo indexado pelas células de origem e destino, com as porcentagens de áreas de interseção calculadas.

3.5.6 2.1-Join-Cells-And-Census.ipynb

Este caderno mescla o arquivo de dados coletados do censo americano às células da grade usando o procedimento descrito em 3.3.2. O arquivo com os dados dos indicadores é indexado pelo *GEOID* (uma sequência de dígitos) do *tract* (setor censitário). No site do US Census é obtido o *shapefile* (Figura 3.8), o arquivo com os dados geoespaciais que pode ser carregado pelo GeoPandas, mesclado com as células da grade por junção espacial e com os indicadores coletados pelo *GEOID*, obtendo todas as interseções *(tract, célula)* possíveis. Essas interseções são agregadas usando o procedimento descrito, resultando em um arquivo com indicadores agregados por célula.

3.5.7 2.2-Join-Cells-And-POI.ipynb

Este caderno realiza agregações de pontos de interesse coletados através do procedimento descrito em 3.3.3. Os pontos coletados já possuem atributos para cada tipo, valendo 0 ou 1 (um ponto pode vir da API classificado em mais de um tipo). Com auxílio do GeoPandas, são obtidos os *buffers* (áreas de alcance) das células (Figura 3.12) e estes são mesclados com os pontos por junção espacial, obtendo-se as combinações *(ponto, buffer)* possíveis. Pontos que intersectam com mais de um buffer são rateados pelas células correspondentes conforme o procedimento descrito, resultando em um arquivo com a contagem de "partes" de pontos por célula.

3.5.8 2.3-Join-All.ipynb

Este caderno é responsável por montar o conjunto de dados final que alimenta o algoritmo de aprendizado de máquina. A identificação dos registros nos arquivos é conferida em busca de repetições nos campos das dimensões espacial ou temporal e, estando tudo correto, os fluxos são carregados em blocos e mesclados aos outros arquivos pelos atributos de identificação comuns.

A opção de processar dados em blocos (usando o parâmetro *chunk* da função *read_csv* do Pandas) foi devido a limitações de memória, pois mesclar grandes conjuntos de dados provoca alto consumo de RAM.

3.5.9 3.1-Random-Forest.ipynb e 3.1-Applying-Boston-Models.ipynb

Estes cadernos usam a saída do caderno 2.3 para treinar e testar modelos por aprendizado de máquina. O caderno *3.1-Random-Forest.ipynb* localiza-se no diretório *boston/ml-models* e realiza testes somente com dados da cidade de Boston, como uma primeira etapa. O caderno *3.1-Applying-Boston-Models.ipynb* localiza-se no diretório *philadelphia/ml-models* e procura extrapolar para outra cidade os modelos gerados a partir do conjunto de dados de Boston. Os resultados obtidos até o momento da entrega do presente Trabalho de Conclusão de Curso são descritos em detalhes no capítulo 4.

Capítulo 4

Resultados

Este trabalho propôs-se a criar um modelo preditivo de mobilidade urbana com dados de Sistemas de Compartilhamento de Bicicletas. As bases de dados abertas de viagens dos serviços *BlueBikes* (Boston, Estados Unidos) e *Indego* (Filadélfia, Estados Unidos) foram processadas e as viagens foram abstraídas em fluxos, conjuntos de viagens agregadas por períodos de tempo e regiões geográficas de origem e destino.

Modelos preditivos foram obtidos por aprendizado de máquina através do algoritmo *Floresta Aleatória* ("Random Forest"). O problema foi modelado como um problema de *regressão*: dado um fluxo de viagens caracterizado por informações socioeconômicas e físicas do espaço e por informações temporais que lhe permitam capturar sua sazonalidade e relação com os dias da semana, períodos do dia e outros, e cuja variável de interesse é o número de viagens agregado para as regiões de origem e destino e o período do tempo, é possível prever esse número de viagens dadas outras regiões e um período de tempo determinado? Também, em caso positivo, quais variáveis mais influenciam a decisão das pessoas de usar a bicicleta em determinados deslocamentos e não em outros?

Inicialmente a modelagem ficou restrita à cidade de Boston, separando-se 80% das amostras de fluxos para aprendizado e 20% para teste do modelo , com resultados animadores para determinados períodos do tempo onde a amostragem é significativa, em especial dias de trabalho nos horários de pico. Em seguida, tentou-se extrapolar o modelo para a cidade da Filadélfia, com todo o conjunto de fluxos de Boston sendo usado para treinamento e o conjunto da Filadélfia usado para teste. Não se chegou a resultados tão animadores quanto no primeiro caso, mas foram formulados questionamentos e obtidos *insights* interessantes sobre como o modelo pode ser melhorado levando em consideração as características de cada cidade, os quais são discutidos neste capítulo.

A modelagem e predição de fluxos de ciclistas do projeto BikeScience, como projeto de pesquisa, deve prosseguir refinando o modelo, agregando novos conjuntos de dados que se considerem interessantes, analisando as variáveis e descartando as que se revelarem pouco significativas, experimentando diferentes algoritmos e bibliotecas de aprendizado de máquina, bem como diferentes técnicas de validação e seleção de hiperparâmetros para modelagem. Tudo que foi descrito neste trabalho representa um estado momentâneo do projeto, tendo sido desenvolvido de maneira incremental e iterativa.

Serão analisados:

- **Acurácia do modelo pelo erro de predição em valor absoluto:** esperam-se predições precisas mas que não indiquem *overfitting* (2.1.3).
- **Correspondência dos quartis de viagens (3.1.3) entre os conjuntos de fluxos reais e preditos:** mesmo que haja divergência de valores, a ideia de que um conjunto pequeno de fluxos concentra a maioria das viagens permite-nos dizer que um modelo é útil se for capaz ao menos de acertar com alguma precisão quais são esses fluxos mais importantes.
- **Importância das características** calculada através da permutação dos valores de cada característica no conjunto de teste e recálculo das predições, tomando-se a diferença entre as acurácias nos dados originais e alterados (Altmann *et al.*, 2010). É introduzida uma característica aleatória (*RANDOM*) para descoberta de características de menos importância (aquelas que se revelarem menos importantes que a aleatória podem ser descartadas).

4.1 Resultados em Boston

4.1.1 Número de viagens e quartil mais significativo

A Figura 4.1 apresenta um gráfico de dispersão onde cada ponto é um fluxo do conjunto de teste, com origem, destino, período e tipo de dia determinados. O eixo horizontal representa os números de viagens reais e o eixo vertical representa as predições. Embora o modelo tenda a prever valores mais baixos, os poucos fluxos com valores mais altos aparecem em sua maioria bem destacados, de forma que parece possível que os mais significativos no conjunto de teste sejam aproximadamente os mesmos para os valores reais e preditos.

Os resultados foram testados para cada mês, período de dia e tipo de dia. Como exemplo, são relatados os resultados para abril de 2019, no período da manhã em dias de trabalho. A

Tabela 4.1. apresenta os quartis de viagens dos fluxos reais e preditos, e as Figuras 4.2 e 4.3 exibem os fluxos do quartil mais significativo, respectivamente os reais e os preditos.

O erro absoluto médio, em número de viagens, é:

- Erro médio: 0, 4049
- Desvio padrão: 1, 7013

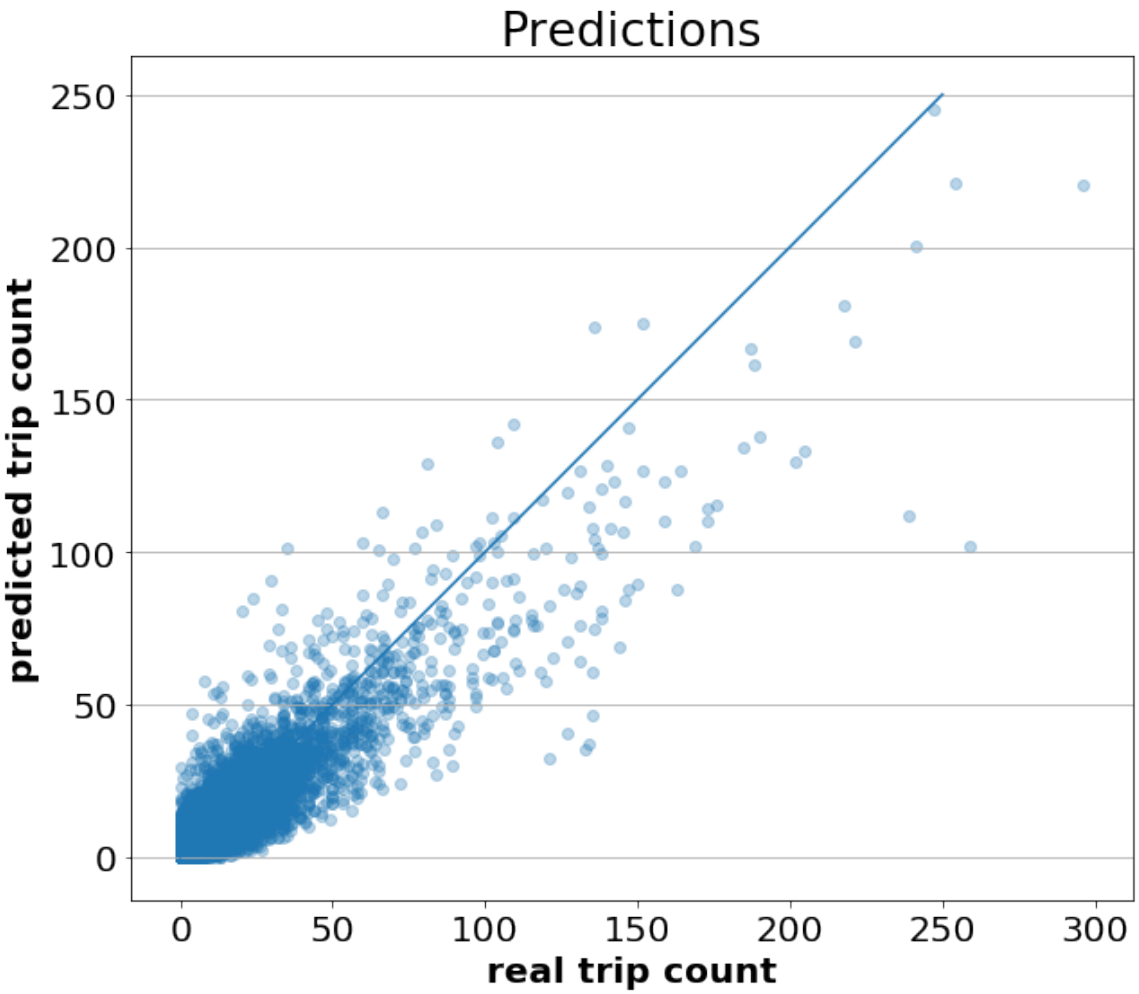

Figura 4.1: *Comparação dos números de viagens reais e preditos para o subconjunto de teste do conjunto de dados do Bluebikes (Boston). Fonte: elaborado pelo autor*

Como o modelo tende a concentrar menos os fluxos, vale a pena olhar mais de perto o excesso de fluxos preditos como significativos. Consideramos um *falso positivo* um fluxo predito como pertencente ao quartil 4 (mais significativo), mas cujo valor real não se encontra nesse quartil. Analogamente, um *falso negativo* é um fluxo realmente significativo mas cujo valor predito não o classificaria como tal. Os falsos positivos podem ainda estar presentes no quartil 3 (o verdadeiro, não o predito pelo modelo), indicando que o modelo os reconhece como significativos, ainda que não tanto quanto são na verdade. As Figuras

	Real				Predito			
Quartil	Mín.	Máx.	Núm. fluxos	% fluxos	Mín.	Máx.	Núm. fluxos	% fluxos
4	50	202	13	0,26%	31	130	22	0,45%
3	17	47	44	0,89%	13	31	64	1,3%
2	7	17	95	1,93%	5	12	152	3,09%
1	1	7	435	8,84%	1	5	636	12,93%

Tabela 4.1: *Separação dos fluxos do conjunto de teste do Bluebikes (Boston) em quartis de viagens para os dias de trabalho e o período da manhã, em abril de 2019*

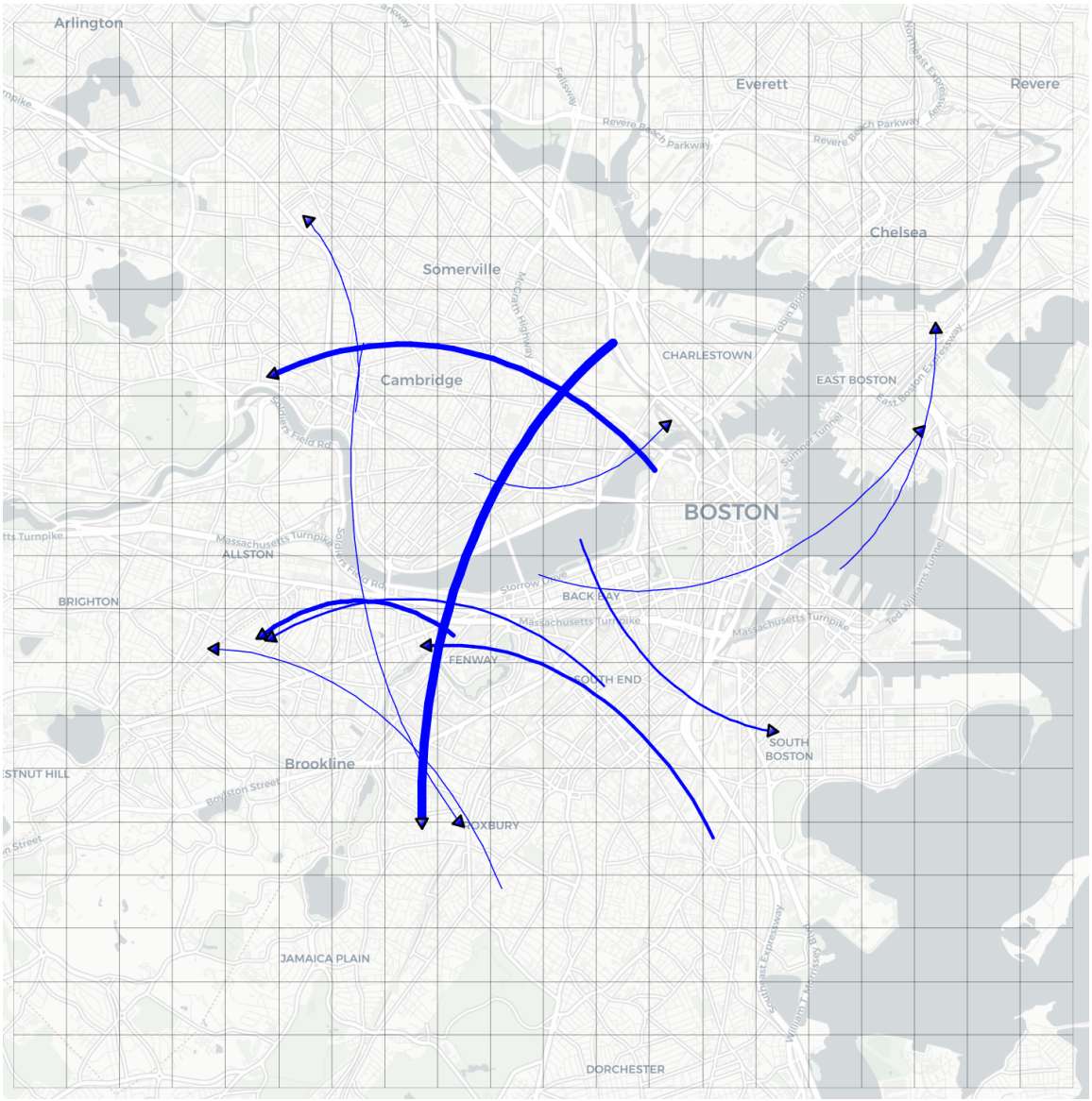

Figura 4.2: *Fluxos do quartil mais significativo real (25% das viagens) do conjunto de teste do Bluebikes (Boston). Fonte: elaborado pelo autor*

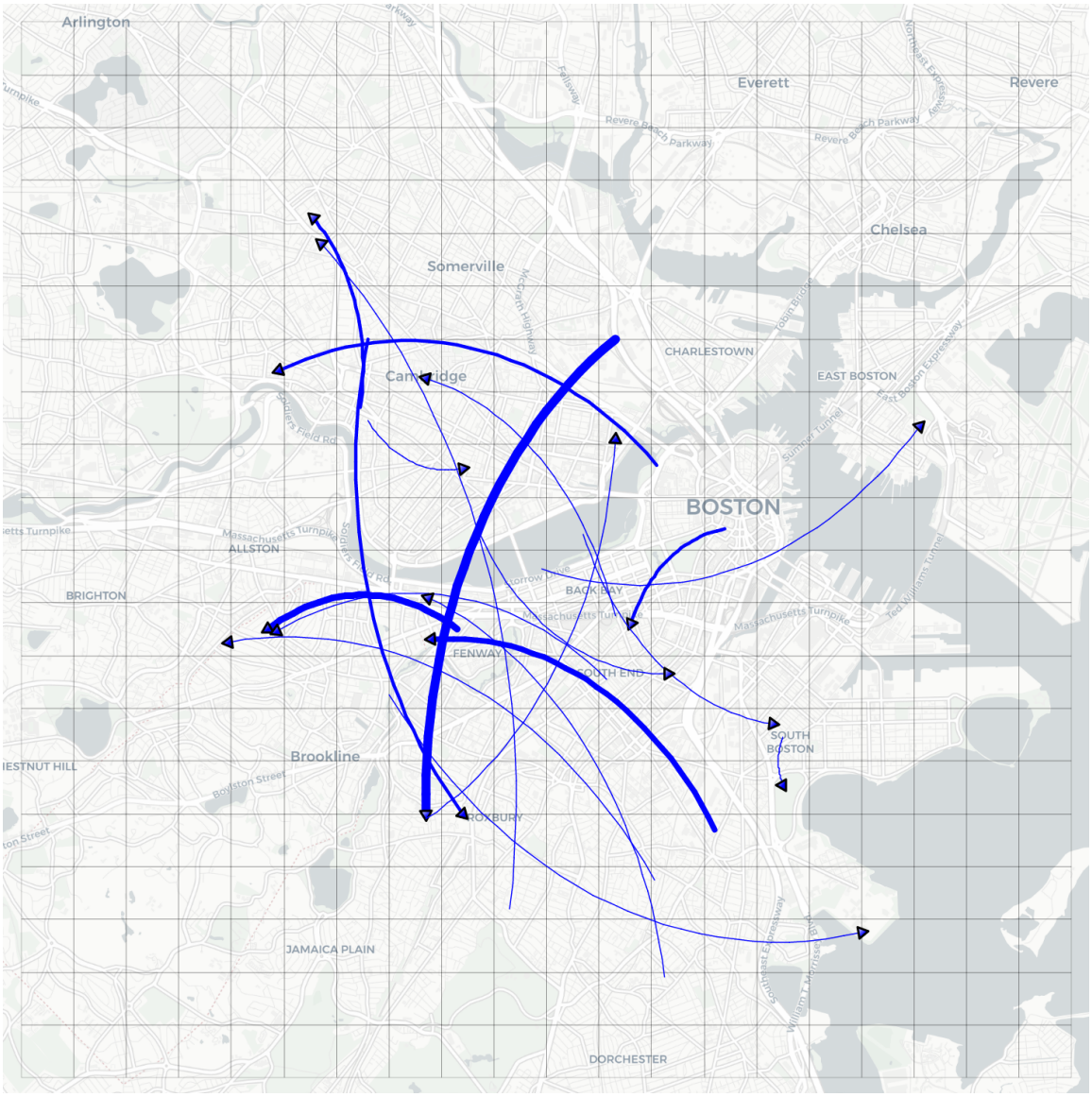

Figura 4.3: *Fluxos do quartil mais significativo predito (25% das viagens) do conjunto de teste do Bluebikes (Boston). Fonte: elaborado pelo autor*

4.4 e 4.5 apresentam, respectivamente, os falsos positivos e negativos, e a presença dos falsos positivos no quartil 3 real.

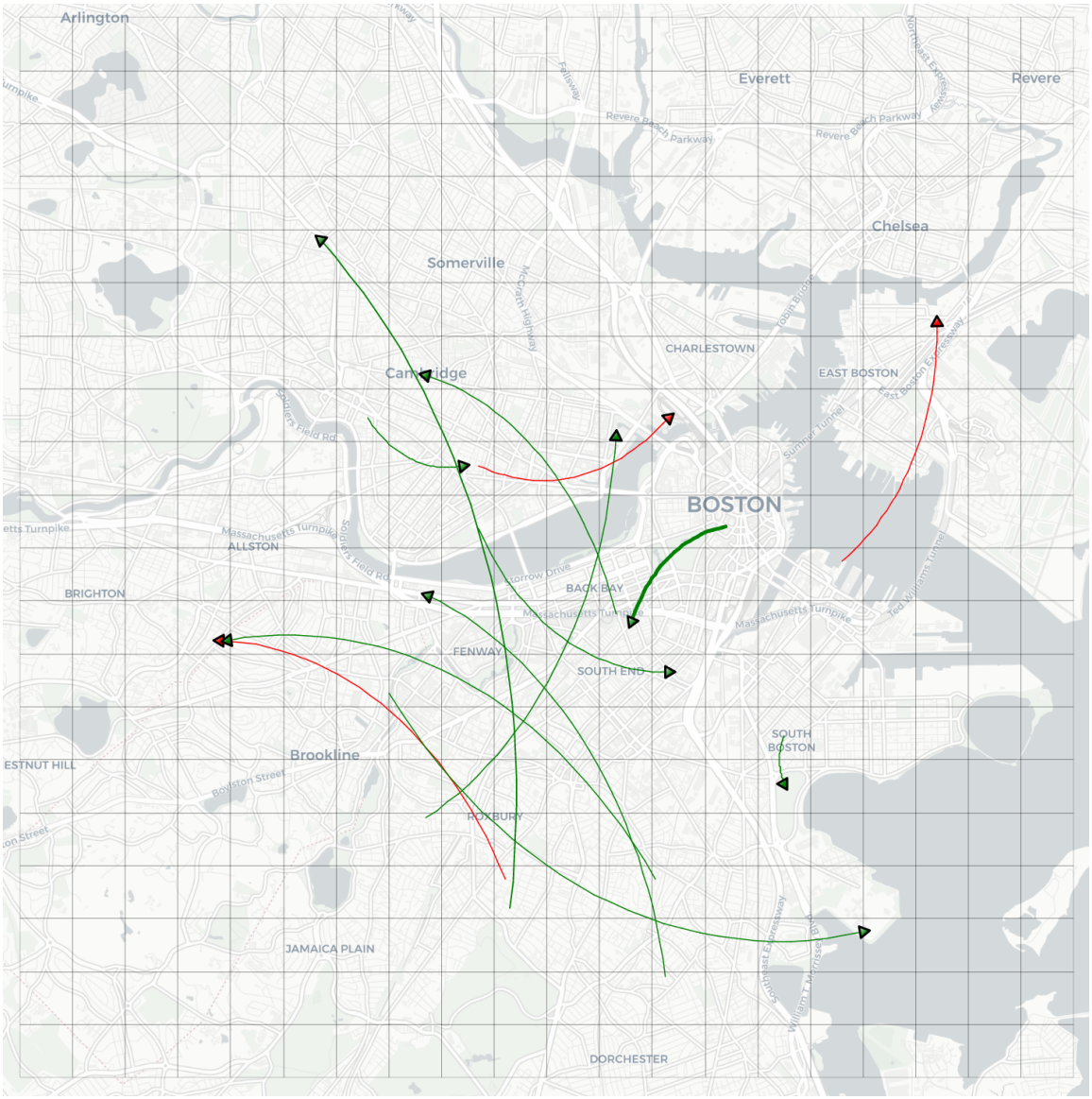

Figura 4.4: *Erros de predição para o quartil mais significativo (25% das viagens) do Bluebikes (Boston). Falsos positivos em verde e falsos negativos em vermelho. Fonte: elaborado pelo autor*

4.1.2 Atributos mais importantes

Tomando-se as características mais significativas pela métrica calculada pelo método da permutação (Altmann *et al.*, 2010), é possível perceber que a separação dos fluxos por período ou tipo de dia, a presença de infraestrutura cicloviária, as condições meteorológicas e de relevo de fato são atributos determinantes para a existência de um fluxo de ciclistas. Os pontos de interesse também aparecem em peso como características significativas. Por outro lado, as características do censo não parecem ter grande influência, mas é interessante

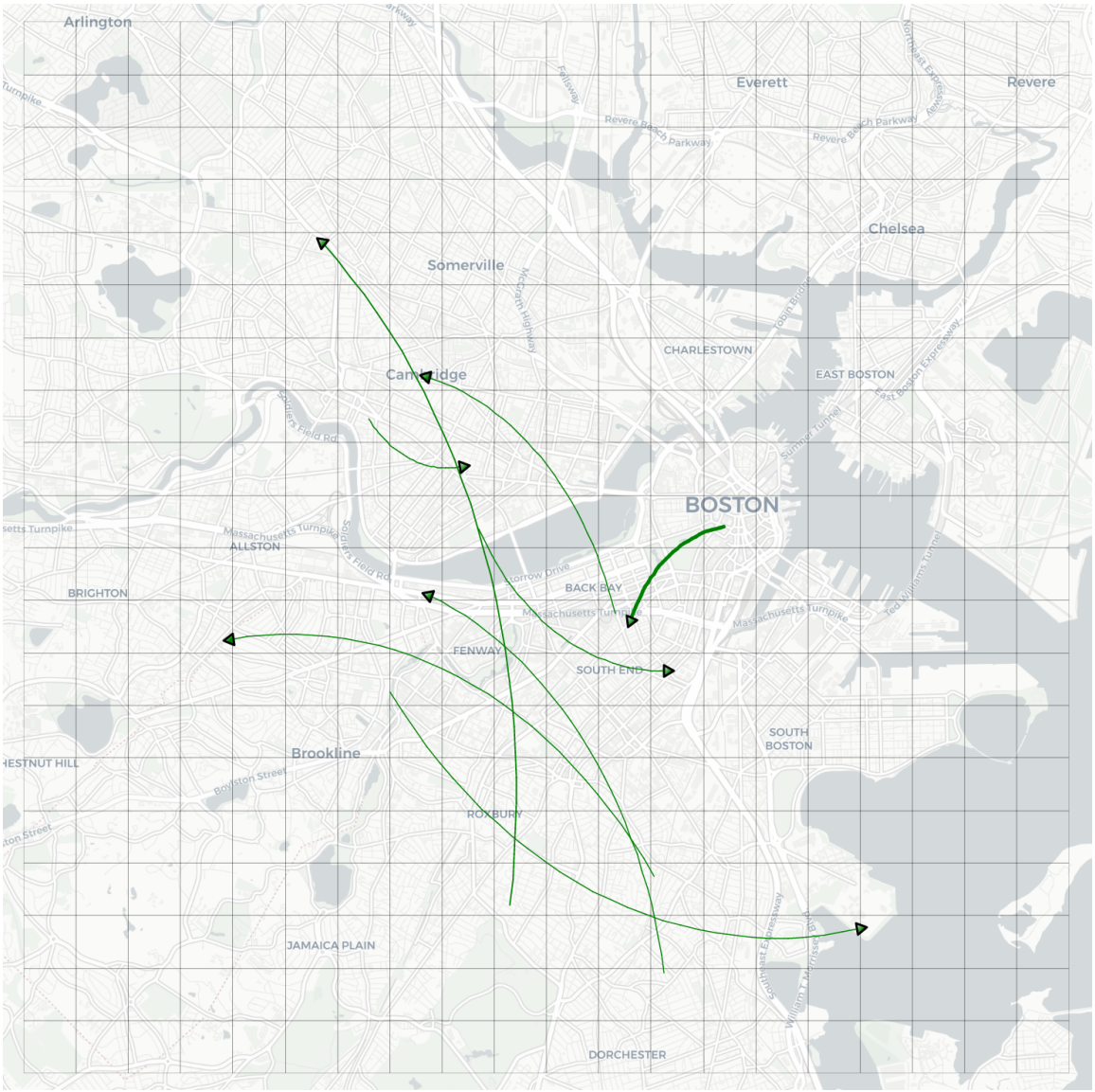

Figura 4.5: *Erros de predição para o quartil mais significativo (25% das viagens) do Bluebikes (Boston) que, na realidade, encontram-se no segundo quartil mais significativo. Fonte: elaborado pelo autor*

a presença do indicador de renda entre as primeiras, sugerindo uma preferência de classe social pelo uso das bicicletas.

A seguir apresentamos as 30 características mais importantes em Boston. A característica aleatória *RANDOM* está na 302ª posição:

1. **distance:** distância percorrida na rota estimada
2. **weekend_or_holiday:** indicador de tipo do dia
3. **lawyer_dest:** ponto de interesse
4. **period:** indicador de período do dia
5. **lawyer_orig:** ponto de interesse
6. **laundry_dest:** ponto de interesse
7. **bikeway_intersect_ratio:** presença de estrutura cicloviária
8. **parking_dest:** ponto de interesse
9. **bar_dest:** ponto de interesse
10. **premise_dest:** ponto de interesse
11. **parking_orig:** ponto de interesse
12. **wspd_mean:** indicador meterorológico
13. **library_dest:** ponto de interesse
14. **real_estate_agency_dest:** ponto de interesse
15. **dew_pt_mean:** indicador meteorológico
16. **university_dest:** ponto de interesse
17. **bar_orig:** ponto de interesse
18. **finance_orig:** ponto de interesse
19. **per_capita_income_last_12_months_mean_orig:** indicador socioeconômico
20. **laundry_orig:** ponto de interesse
21. **library_orig:** ponto de interesse
22. **rh_mean:** indicador meteorológico
23. **average_elevation_dest:** indicador de altitude

24. **finance_dest:** ponto de interesse
25. **feels_like_std:** indicador meteorológico
26. **female_18_and_19_years_mean_dest:** indicador socioeconômico
27. **heat_index_max:** indicador meteorológico
28. **feels_like_max:** indicador meteorológico
29. **female_college_less_than_1_year_mean_dest:** indicador socioeconômico
30. **bank_dest:** ponto de interesse

4.1.3 Discussão

Treinar e validar um modelo preditivo usando dados de uma única cidade pode não parecer tão válido ou útil, afinal, cada cidade possui características diferentes. De fato, ao se limitar demais o modelo, realizando treinamentos separados para períodos e tipos de dia diferentes, conseguem-se resultados muito próximos do ótimo, em especial para manhãs e finais de tarde de dias de trabalho, quando a amostragem é mais significativa. Esse resultado, apesar de ter impressionado, pode ser um caso claro de "overfitting".

A opção de limitar-se inicialmente a uma cidade permitiu testar ideias e verificar a viabilidade de um modelo preditivo de fluxos de mobilidade. Como exemplo, uma substancial melhora na acurácia foi obtida ao se decidir distribuir os pontos de interesse de maneira fracionada pelas células.

4.2 Modelo de Boston extrapolado para Filadélfia

Tendo-se conseguido um bom resultado para uma cidade como Boston, testamos a modelagem com dados de mais cidades, e assim começamos a identificar características que diferenciam as cidades e formas de modelá-las. Inicialmente, testamos a capacidade de generalização do modelo obtido com os dados do Bluebikes, usando os dados do Indego como conjunto de teste.

Embora não se esperasse uma acurácia tão grande quanto o modelo em 4.1, o objetivo é procurar por pontos de melhoria que possam efetivamente aumentar a capacidade de generalização. A primeira melhoria encontrada foi, antes de aplicar o algoritmo de aprendizado, realizar a normalização dos dados de entrada (2.1.2) para ambas as cidades, colocando-os na mesma escala através do cálculo da estatística Z (normal padrão).

4.2.1 Número de viagens e quartil mais significativo

A Figura 4.6 revela que o modelo está excessivamente ajustado (*overfitting*) nas características de Boston e, portanto, ainda é preciso procurar formas de modelar as diferenças entre as duas cidades. As Figuras 4.7, 4.8, 4.9 e 4.10 mostram como está a predição do quartil mais significativo, do mesmo modo como foi feito somente com os dados de Boston (seção 4.1).

O erro absoluto médio, em número de viagens, é:

- Erro médio: 1, 5863
- Desvio padrão: 4.1048

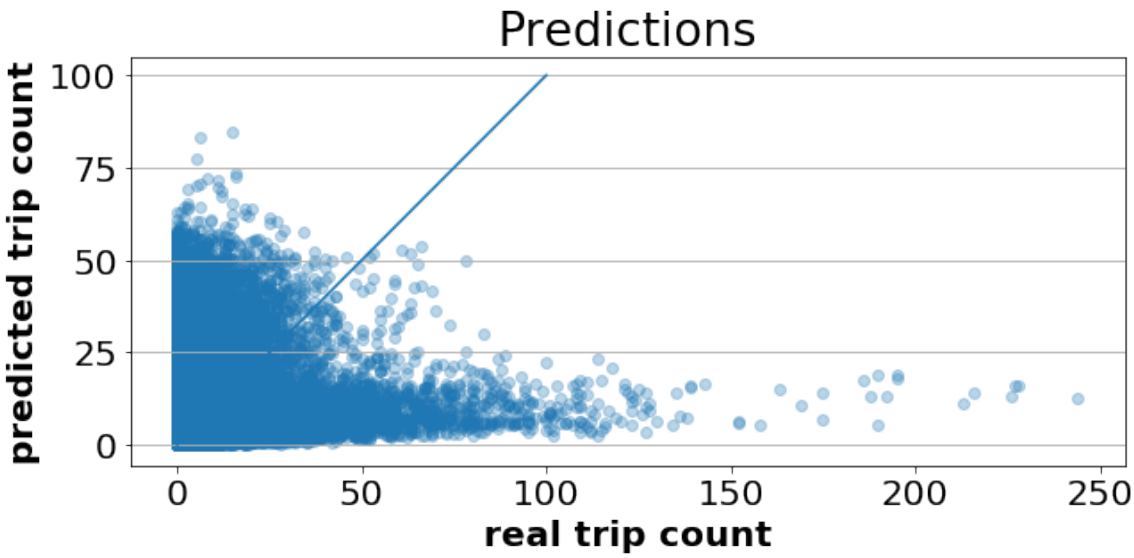

Figura 4.6: *Comparação dos números de viagens reais e preditos para o conjunto de dados do Indego (Filadélfia). Predição realizada a partir do modelo obtido com o Bluebikes (Boston). Fonte: elaborado pelo autor*

	Real				Predito			
Quartil	Mín.	Máx.	Núm. fluxos	% fluxos	Mín.	Máx.	Núm. fluxos	% fluxos
4	32	106	38	0,67%	12	41	203	3,56%
3	16	32	84	1,47%	7	12	438	7,68%
2	7	16	172	3,02%	4	7	765	13,42%
1	1	7	809	14,19%	1	4	2235	32,21%

Tabela 4.2: *Separação dos fluxos do conjunto de dados do Indego (Filadélfia) em quartis de viagens para os dias de trabalho e o período da manhã, em abril de 2019. Predições realizadas a partir do modelo obtido com o Bluebikes (Boston).*

Como este modelo apresenta muitos erros, é importante analisar com cuidado os fluxos mais importantes, tanto reais quanto preditos. A Figura 4.6 revela que os fluxos com alto

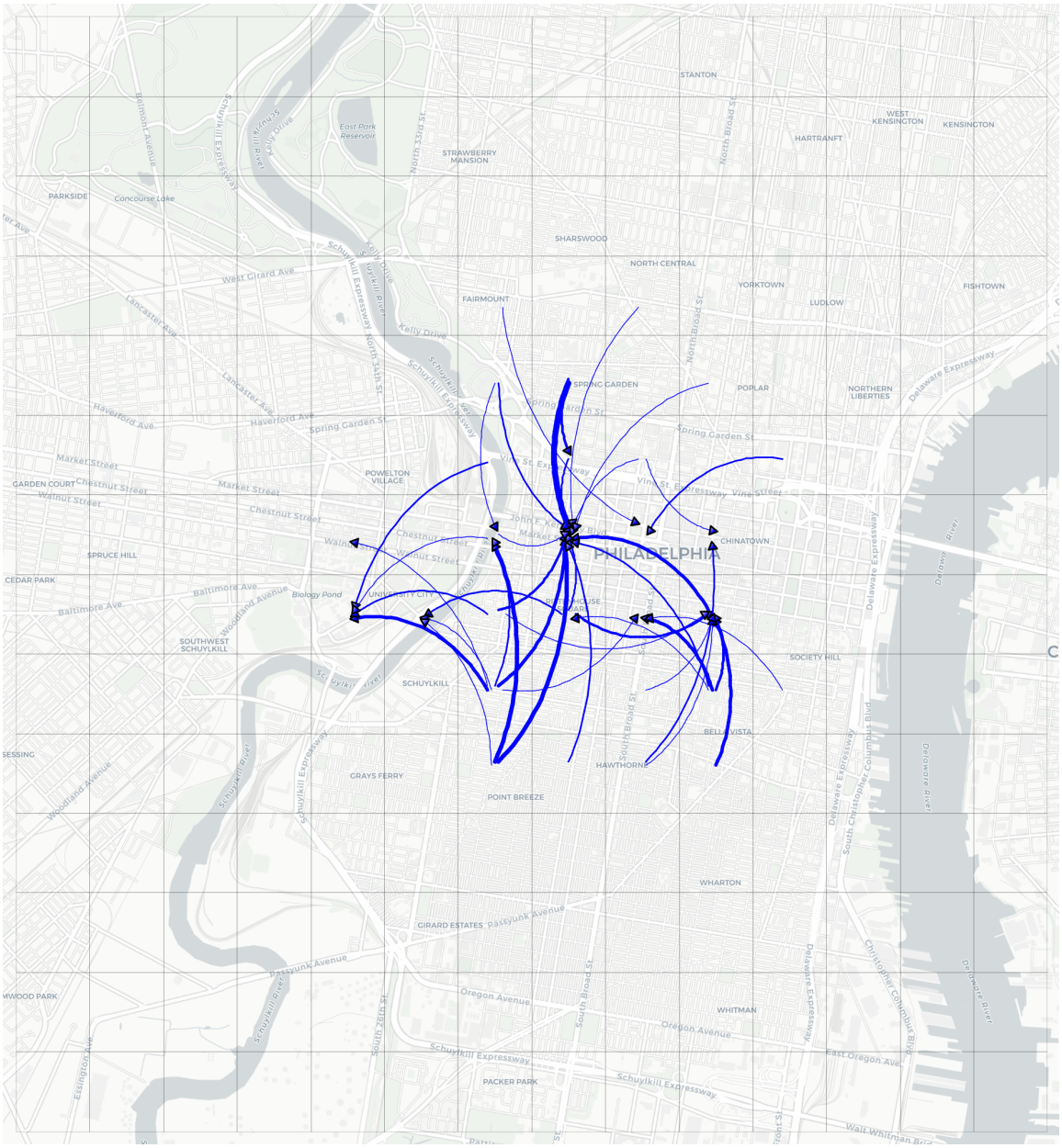

Figura 4.7: *Fluxos reais do quartil mais significativo (25% das viagens) do conjunto de dados do Indego (Filadélfia). Predição realizada a partir do modelo obtido com o Bluebikes (Boston). Fonte: elaborado pelo autor*

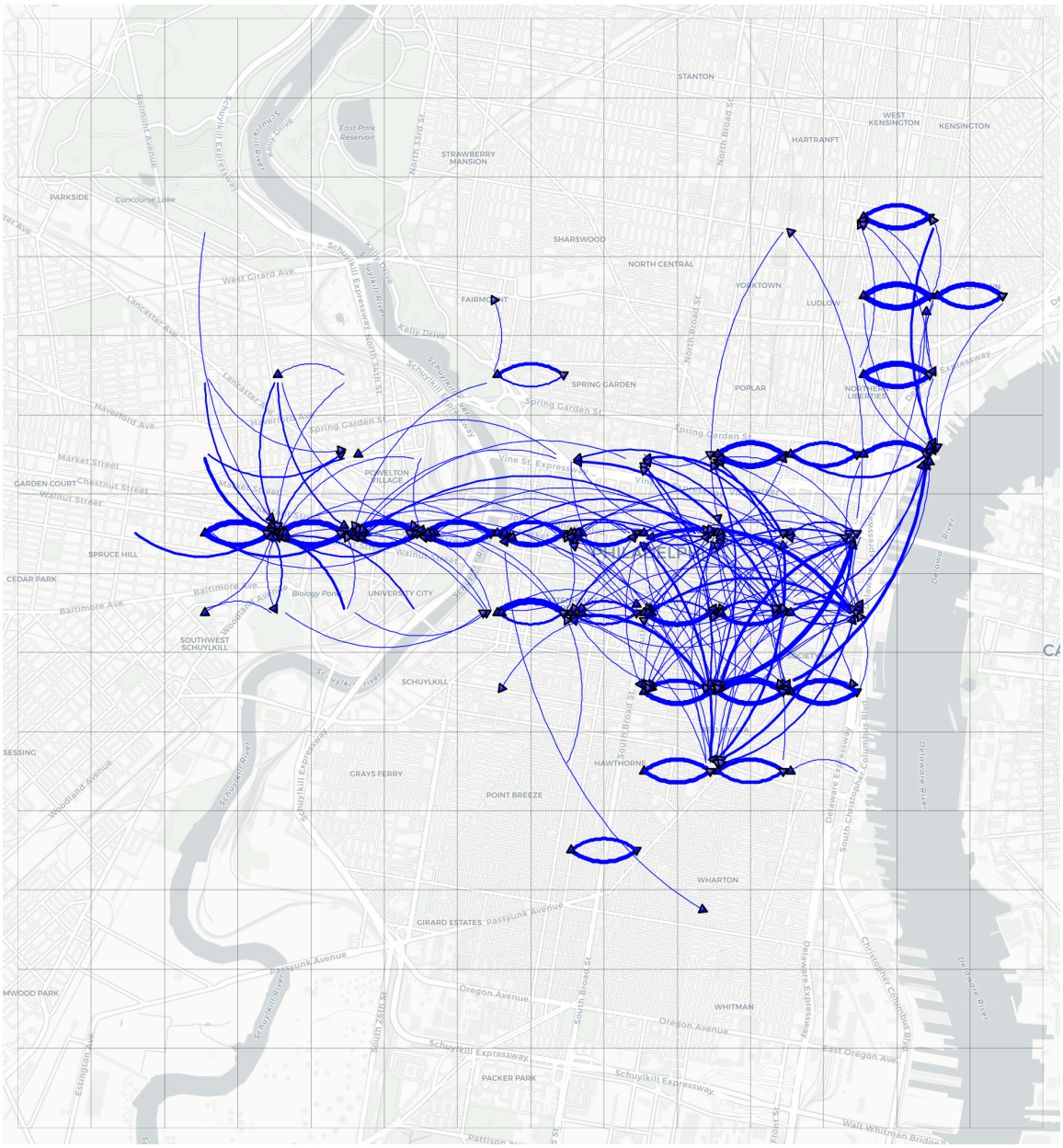

Figura 4.8: *Fluxos preditos para o quartil mais significativo (25% das viagens) do conjunto de dados do Indego (Filadélfia). Predição realizada a partir do modelo obtido com o Bluebikes (Boston). Fonte: elaborado pelo autor*

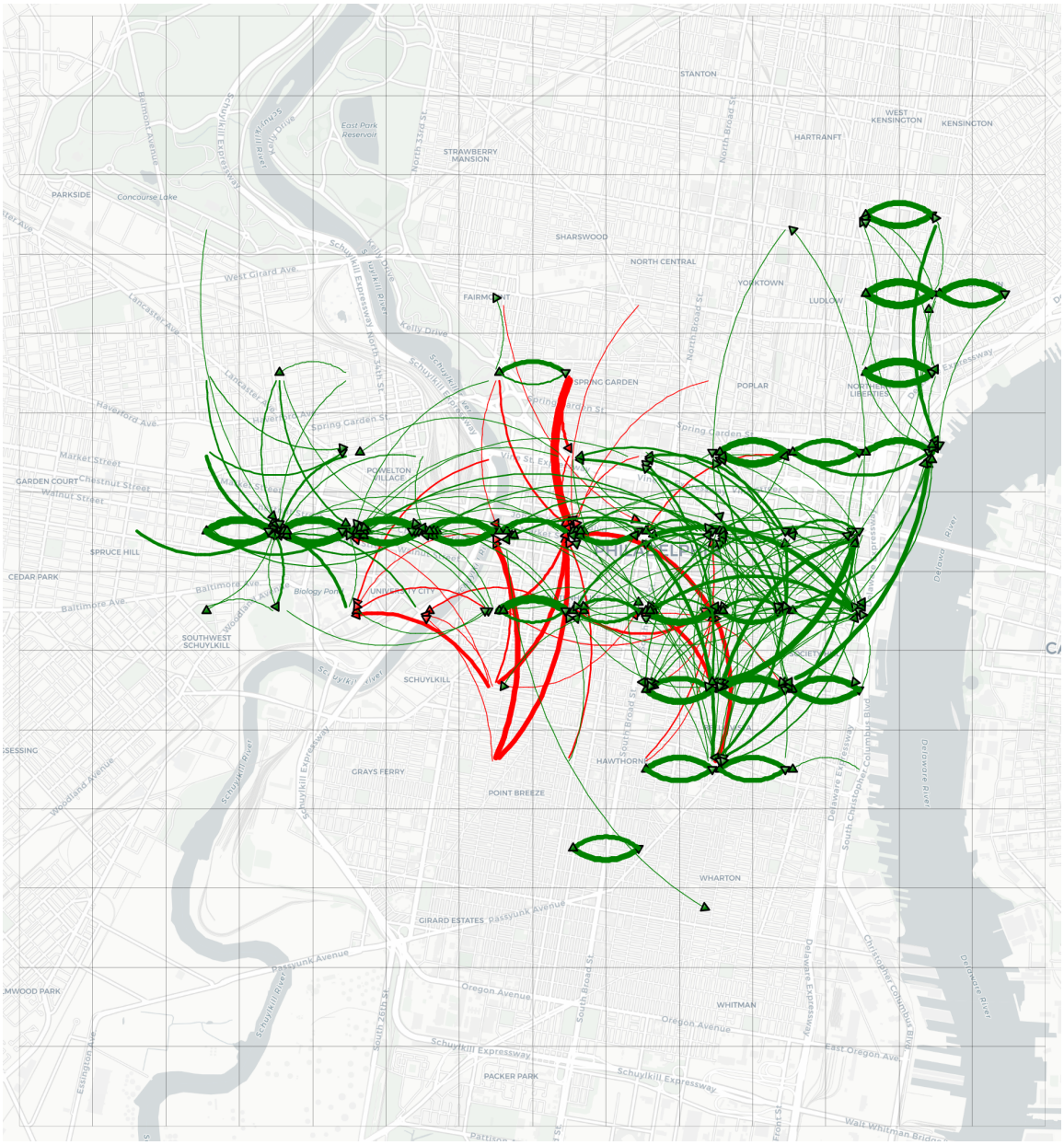

Figura 4.9: *Erros de predição para o quartil mais significativo (25% das viagens) do Indego (Filadélfia). Falsos positivos em verde e falsos negativos em vermelho. Fonte: elaborado pelo autor*

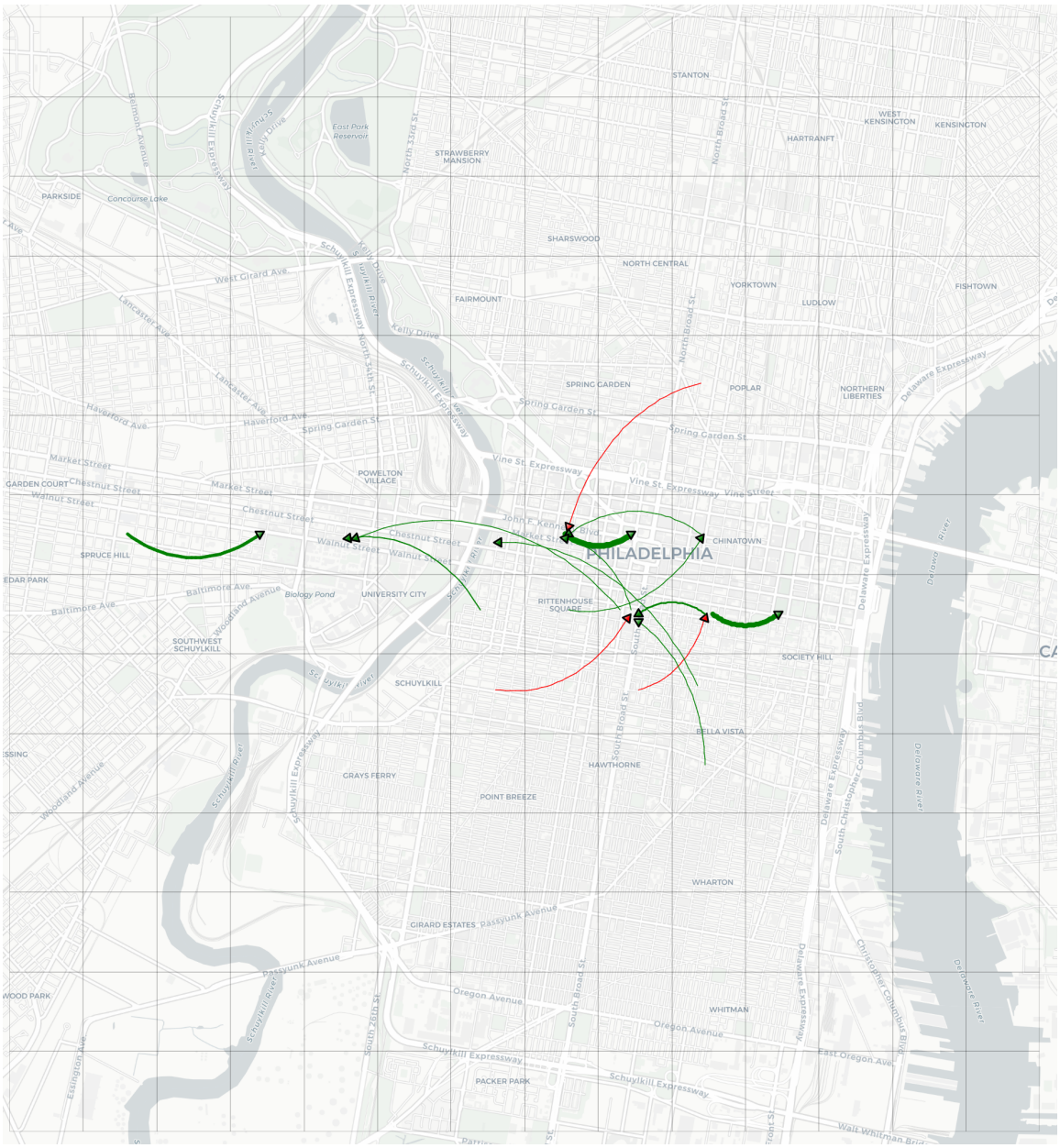

Figura 4.10: *Erros de predição para o quartil mais significativo (25% das viagens) do Indego (Filadélfia) que, na realidade, encontram-se no segundo quartil mais significativo. Fonte: elaborado pelo autor*

número real de viagens são preditos como tendo baixo número; e os fluxos preditos como tendo alto número de viagens são em sua maioria fluxos com baixo número. Para esta análise, diferentemente do que foi feito para Boston em separado (seção 4.1), definimos os *falsos positivos* e os *falsos negativos* através de cortes nos valores: os falsos positivos são os fluxos com até 40 viagens reais e a partir de 30 viagens preditas; os falsos negativos são os fluxos com no mínimo 100 viagens preditas, todos eles preditos como tendo menos que 25 viagens. Esses fluxos podem ser vistos nas Figuras 4.11 e 4.12, e suas localizações podem ser inspecionadas em detalhes em busca de suas características. Por exemplo, os falsos negativos revelam uma célula que é destino de vários fluxos, na qual se localiza a *Commerce Street*, o que leva imediatamente à pergunta: os pontos de interesse estão sendo capturados adequadamente?

4.2.2 Atributos mais importantes

Não há sentido em se realizar uma análise das características deste modelo, visto que os dados de treinamento são os mesmos do anterior. O que se fez foi gerar um modelo separado para Filadélfia e aferir daí suas características mais influentes.

A característica aleatória *RANDOM* aparece na 826ª posição, indicando que todo o conjunto de features possui importância nessa cidade. De fato, ao contrário de Boston, os indicadores socioeconômicos (obtidos do censo americano) aparecem em grande número no topo da lista. Parece que os profissionais de nível técnico têm alguma predileção especial (ou necessidade) pelo uso das bicicletas!

A seguir apresentamos as 30 características mais relevantes:

1. **distance:** distância percorrida na rota estimada
2. **weekend_or_holiday:** indicador de tipo do dia
3. **period:** indicador de período do dia
4. **female_professional_school_degree_mean_dest:** indicador socioeconômico
5. **female_professional_school_degree_max_orig:** indicador socioeconômico
6. **male_professional_school_degree_mean_orig:** indicador socioeconômico
7. **wspd_mean:** indicador meteorológico
8. **bikeway_intersect_ratio:** presença de infraestrutura cicloviária
9. **male_professional_school_degree_mean_dest:** indicador socioeconômico
10. **female_professional_school_degree_mean_orig:** indicador socioeconômico

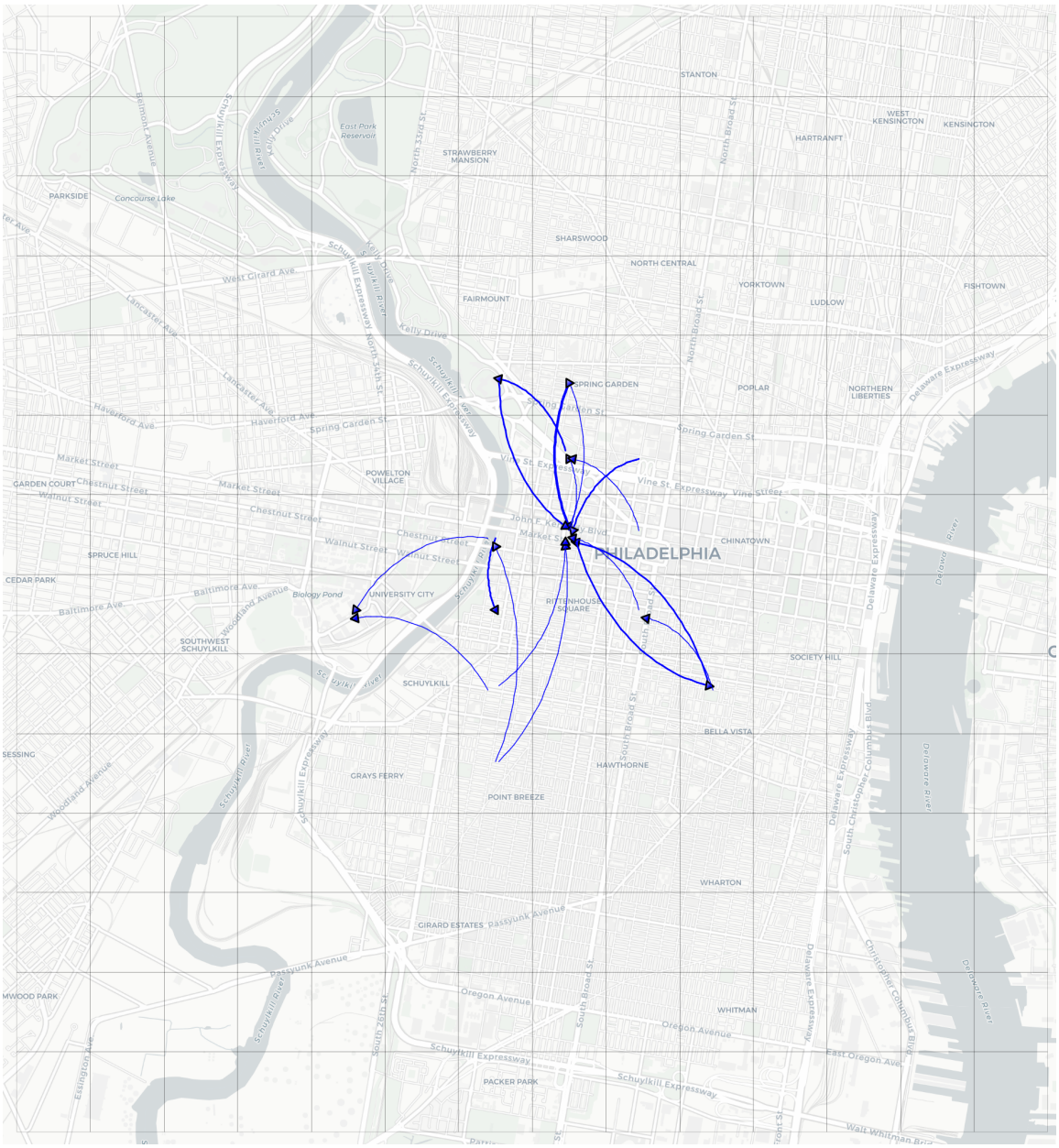

Figura 4.11: *Fluxos mais significativos do Indego (Filadélfia) preditos como não significativos. Predição realizada a partir do modelo obtido com o Bluebikes (Boston). Fonte: elaborado pelo autor*

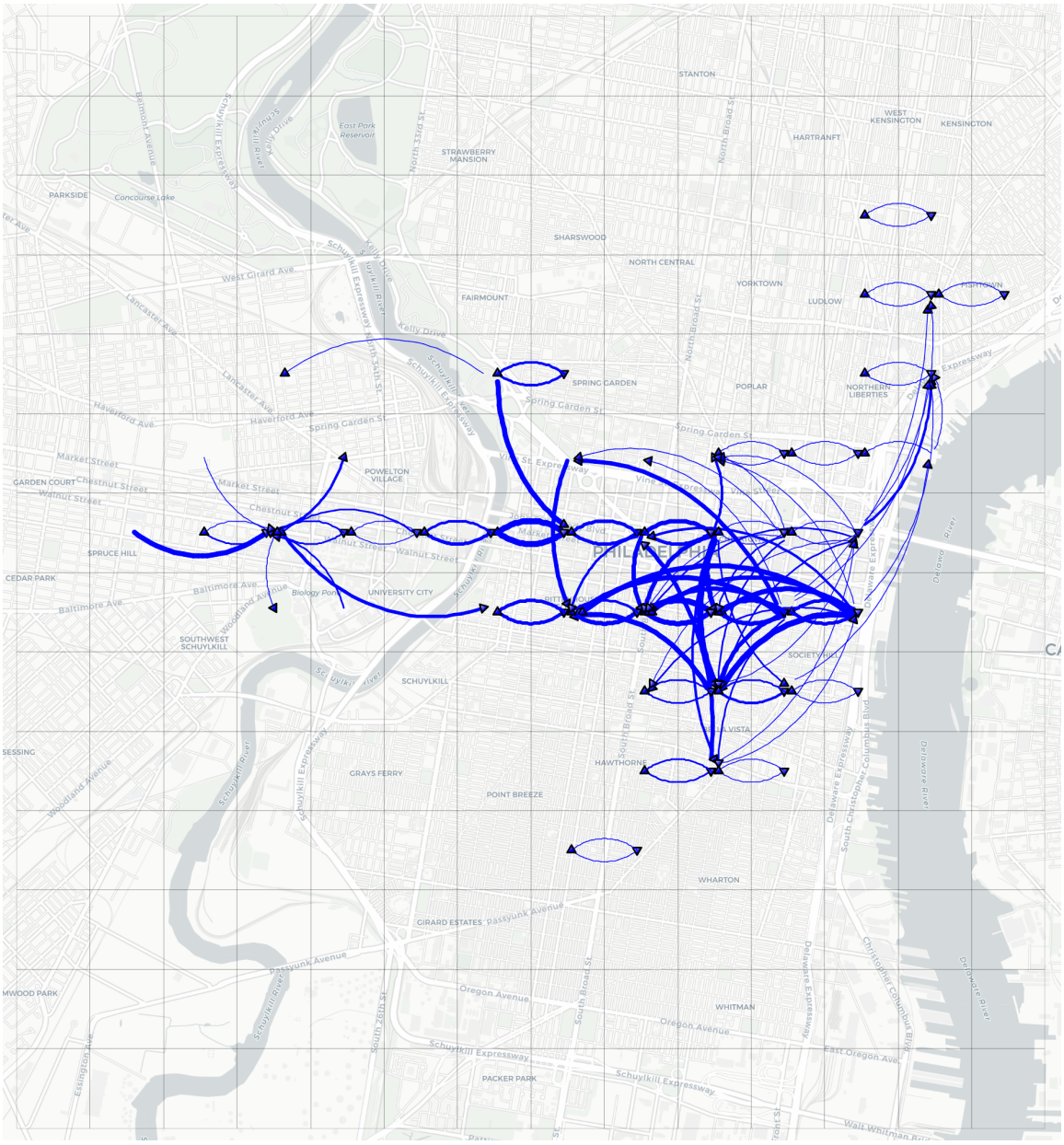

Figura 4.12: *Fluxos menos significativos do Indego (Filadélfia) preditos como mais significativos. Predição realizada a partir do modelo obtido com o Bluebikes (Boston). Fonte: elaborado pelo autor*

11. **lodging_dest:** ponto de interesse
12. **restaurant_dest:** ponto de interesse
13. **bar_dest:** ponto de interesse
14. **female_10th_grade_std_dest:** indicador socioeconômico
15. **lawyer_orig:** ponto de interesse
16. **feels_like_mean:** indicador meteorológico
17. **doctor_orig:** ponto de interesse
18. **wc_mean:** indicador meteorológico
19. **female_10_to_14_years_mean_dest:** indicador socioeconômico
20. **wc_max:** indicador meteorológico
21. **temp_max:** indicador meteorológico
22. **rh_mean:** indicador meteorológico
23. **lawyer_dest:** ponto de interesse
24. **insurance_agency_orig:** ponto de interesse
25. **real_estate_agency_dest:** ponto de interesse
26. **hardware_store_dest:** ponto de interesse
27. **book_store_orig:** ponto de interesse
28. **bank_orig:** ponto de interesse
29. **doctor_dest:** ponto de interesse
30. **per_capita_income_last_12_months_mean_orig:** indicador socioeconômico

4.2.3 Discussão

Ao se aplicar sobre o Indego um modelo treinado com dados de viagens do Bluebikes, não há a pretensão de que esse modelo tenha alta capacidade de generalização logo de início, com somente uma cidade usada para aprendizado. No entanto, esse modelo serve como teste para ideias que possam melhorar essa capacidade de generalização.

Os mapas desta seção apresentam os resultados da *segunda* tentativa de modelagem; a primeira tentativa apresentou resultados muito mais imprecisos. O que provocou essa melhora foi a ideia de colocar os valores das características na mesma escala através de

normalização (cálculo da estatística Z), pois um valor que é “alto” em Boston pode ser “baixo” na Filadélfia e vice-versa.

4.3 Modelo conjunto de Boston e Filadélfia

Uma modelagem mais robusta deve ser obtida a partir de dados de mais cidades. O esperado para o longo prazo, conforme o projeto de pesquisa evoluir, é que se tenha um modelo aprendido a partir de mais cidades com características diferentes, e que esse modelo possa acertar com boa precisão os fluxos de viagens de novas cidades não visitadas no treinamento, para as quais os dados estariam disponíveis para averiguação. Assim, ao se aplicá-lo a uma cidade onde um Sistema de Compartilhamento de Bicicletas é inexistente, haverá a confiança de que os fluxos preditos indicarão onde uma infraestrutura deve ser construída.

4.3.1 Número de viagens e quartil mais significativo

É uma boa notícia que a Figura 4.13 revele novamente um bom encaixe do conjunto de teste (20% de toda a amostra) para um modelo treinado com duas cidades com características diferentes. É possível perceber que Filadélfia apresenta menos viagens, pois a maioria de seus fluxos mais volumosos encontram-se abaixo de 100 viagens, valor que é superado por diversos fluxos em Boston.

O erro absoluto médio, em número de viagens, é:

- Erro médio: $0,4577$
- Desvio padrão: $1,6122$

4.3.2 Atributos mais importantes

Uma análise acurada da importância das caracteristias neste modelo revela o que as cidades têm em comum. A distância, a presença da estrutura cicloviária e de pontos de interesse, os indicadores meteorológicos e a dimensão temporal revelam-se fundamental, embora não suficiente, para caracterizar os fluxos. Um atrator de viagens revelado tanto nos mapas quanto no *ranking* de atributos é a presença de universidades.

A característica aleatória *RANDOM* foi classificada na 648ª posição. As 30 características mais importantes foram:

1. **distance:** distância percorrida na rota estimada
2. **weekend_or_holiday:** indicador de tipo do dia

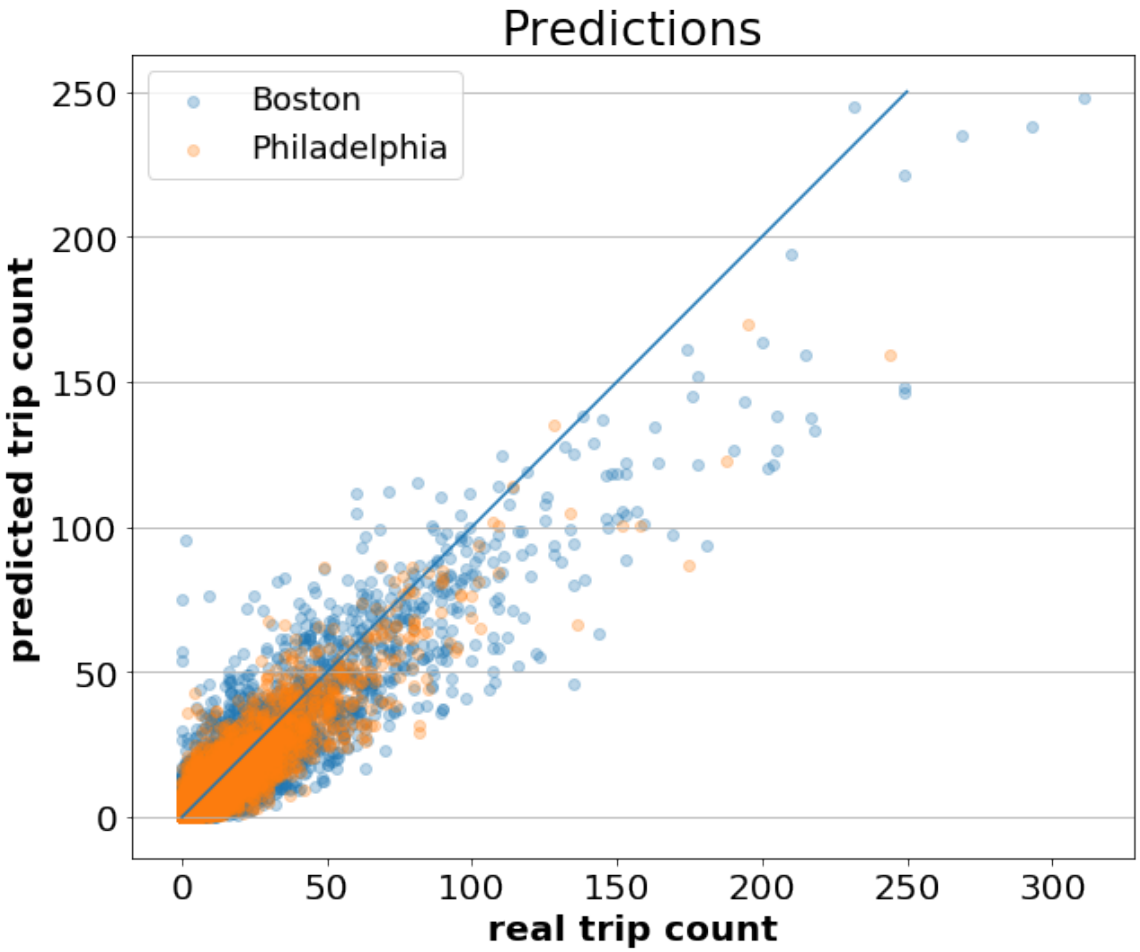

Figura 4.13: *Comparação dos números de viagens reais e preditos para o subconjunto de teste da união dos conjunto de viagens do Bluebikes (Boston) e do Indego (Filadélfia). Fonte: elaborado pelo autor*

3. **lawyer_dest:** ponto de interesse
4. **period:** indicador de período do dia
5. **lawyer_orig:** ponto de interesse
6. **laundry_dest:** ponto de interesse
7. **bikeway_intersect_ratio:** presença de estrutura cicloviária
8. **parking_dest:** ponto de interesse
9. **bar_dest:** ponto de interesse
10. **premise_dest:** ponto de interesse
11. **parking_orig:** ponto de interesse
12. **wspd_mean:** indicador meterorológico

13. **library_dest:** ponto de interesse
14. **real_estate_agency_dest:** ponto de interesse
15. **dew_pt_mean:** indicador meterorológico
16. **university_dest:** ponto de interesse
17. **bar_orig:** ponto de interesse
18. **finance_orig:** ponto de interesse
19. **per_capita_income_last_12_months_mean_orig:** indicador socioeconômico
20. **laundry_orig:** ponto de interesse
21. **library_orig:** ponto de interesse
22. **rh_mean:** indicador meterorológico
23. **average_elevation_dest:** indicador de altitude
24. **finance_dest:** ponto de interesse
25. **feels_like_std:** indicador meterorológico
26. **female_18_and_19_years_mean_dest:** indicador socioeconômico
27. **heat_index_max:** indicador meterorológico
28. **feels_like_max:** indicador meterorológico
29. **female_college_less_than_1_year_mean_dest:** indicador socioeconômico
30. **bank_dest:** ponto de interesse

4.3.3 Discussão

Os mesmos mapas das seções anteriores foram graficados para o modelo conjunto, usando o subconjunto de fluxos de teste. Revelou-se que o modelo consegue prever vários fluxos importantes (aqueles no quartil 4), porém um pouco abaixo da precisão obtida usando-se somente a cidade de Boston.

Essa é o regressor que deverá eventualmente ser colocado em produção, mas para isso muito deve ser feito ainda. É possível, por exemplo, eliminar características consideradas irrelevantes para economizar tempo de processamento e memória, obter novos conjuntos de dados que possam caracterizar mais os fluxos e também as cidades onde esses fluxos se encontram, aplicar novas ideias baseadas na teoria de aprendizado de máquina que se revelarem úteis, entre outras coisas.

Capítulo 5

Conclusão

Nos últimos anos, o aprendizado de máquina e a ciência de dados tornaram-se termos da moda, alavancados pelo fenômeno do *big data* e o uso do *deep learning* com sucesso em processamento e classificação de imagens. Este trabalho foi a oportunidade do autor tomar contato com esse tipo de tecnologia e encarar seus pontos fortes, suas limitações, aspectos teóricos e técnicos, inclusive as dificuldades.

Apesar do (*"hype"*) em torno do conceito, o aprendizado de máquina é poderoso e útil mas não é panaceia; em especial quando se trata de um trabalho de pesquisa que procura descobrir a viabilidade de algo ainda não tentado. O projeto de pesquisa do *BikeScience* está em andamento; muito se avançou porém muito ainda há que ser feito. Este trabalho não pretende, nem de longe, dar por resolvido o problema da modelagem de fluxos de ciclistas, porém espera contribuir com um arcabouço computacional consistente e com o teste da relevância (ou não) de determinados conjuntos de dados para a acurácia de um modelo desse tipo.

A análise exploratória prévia dos dados revelou-se de grande valor: classificar os fluxos de ciclistas conforme ocorrem em dias de trabalho ou finais de semana, em horários diferentes ou mesmo por mês, capturando a variação sazonal, gerou um conjunto de atributos considerados como de alta relevância pelos métodos de análise utilizados. Também, os conjuntos de dados coletados – indicadores meteorológicos, pontos de interesse, relevo e altitude e outros – foram em sua maioria significativos e a maioria dos atributos obtidos a partir deles sem dúvida deverão permanecer na modelagem. Já os indicadores socioeconômicos, sobre os quais um dos questionamentos iniciais era quanto à sua importância para esta modelagem, provou-se de importância variável: são influentes na Filadélfia porém pouco significativos em Boston. No entanto, o indicador de renda mensal média nas regiões analisadas ficou bem classificado pelo método das permutações.

O avanço na qualidade dos modelos deu-se aos saltos: tentativas de incorporar novos dados e formas de calcular atributos são tentadas sucessivamente, sem avanços significativos, até o momento em que uma ideia prova-se funcional. Como exemplo:

1. Gerar modelos para diferentes subconjuntos dos dados, como período do dia (manhã, horário de almoço e final da tarde) e tipo (dia de trabalho e finais de semana): como essas divisões são muito bem marcadas, em especial os modelos de horários de pico (manhã e final da tarde de dias de trabalho), elas produzem modelos focados na distribuição particular de cada subconjunto dos dados com maior precisão.

2. Distribuir os pontos de interesse e as regiões socioeconômicas proporcionalmente pelas células ao redor: como a grade é arbitrária, isso reduziu o efeito nas bordas das células. Viagens que partem de ou chegam a uma estação próxima às bordas ou cantos, podem ser relativas a pontos de interesse próximos em uma célula vizinha.

3. Colocar dados de duas cidades na mesma escala antes de aplicar o algoritmo de aprendizado: cada cidade possui uma distribuição diferente dos valores dos atributos. Se os valores forem normalizados, o algoritmo poderia capturar as diferenças entre as cidades de forma mais precisa.

Esses itens evidenciam a importância da análise exploratória dos dados (1), do conhecimento do negócio, ou seja, do funcionamento de um Sistema de Compartilhamento de Bicicletas (2) e da aplicação do arcabouço teórico (3).

A pesquisa realizada é original no sentido de que um fluxo de mobilidade leva em consideração as origens das viagens, não apenas os destinos. Estes já foram objetos de modelagem em estudos prévios, de detecção dos *hubs* ou regiões atratoras de viagens de bicicletas. Levar em conta as origens permite capturar a natureza pendular de muitos dos fluxos mais significativos, isto é, se as pessoas vão para determinados locais em algum horário do dia, elas fazem o trajeto de volta em outro horário.

As dificuldades encontradas ao longo do caminho permitem encarar as limitações desse tipo de modelagem. A análise com mais de uma cidade está em seu início e muito há que ser trabalhado ainda. Boston e Filadélfia possuem padrões de mobilidade muito diferentes, e determinar em um mesmo modelo o que é importante em cada uma é um desafio. Também, a arbitrariedade da grade exige soluções criativas ao modelar as características de uma célula, levando em conta as áreas vizinhas. As ferramentas utilizadas também possuem limitações por trabalharem em memória RAM e não serem escaláveis sem artifícios complexos, tendo sido necessária a alocação de máquinas cada vez mais poderosas e com mais memória para realizar junções em conjuntos de dados com centenas de colunas e milhões de linhas.

Como sugestão de trabalhos futuros, já se tem algumas ideias para melhorar o modelo atual, no estado em que é descrito neste Trabalho:

- A partir da análise da importância das features, descartar as que se mostraram pouco relevantes.
- Testar mais conjuntos de dados que podem influir no tráfego de bicicletas. Como exemplo, mas não se limitando a, dados de acidentes, criminalidade, zoneamento ou ocupação territorial.
- Modelar as regiões de origem e destino como hexágonos ao invés de retângulos. Os cantos das células, como estão, são áreas cuja modelagem é mais imprecisa.
- Modelar características das cidades, como forma de tentar ensinar o modelo a diferenciá-las. Como exemplo, dados econômicos e populacionais agregados para a cidade.

Participar do projeto BikeScience foi um prazer pessoal e uma fonte imensa de aprendizado. Ficam os votos de que a pesquisa seja um sucesso e o desejo de boa sorte a quem for continuar o trabalho.

Referências

[Altmann *et al.* 2010] André Altmann, Laura Toloşi, Oliver Sander e Thomas Lengauer. "Permutation importance: a corrected feature importance measure". Em: *Bioinformatics* 26.10 (abr. de 2010), pgs. 1340–1347. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq134. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/26/10/1340/16892402/btq134.pdf. URL: https://doi.org/10.1093/bioinformatics/btq134 (citado nas pgs. 54, 58).

[Bureau 2018] United States Census Bureau. *American Community Survey 5-Year Data (2009-2017)*. 2018. URL: https://www.census.gov/data/developers/data-sets/acs-5year.html (citado na pg. 34).

[Bureau 2019] United States Census Bureau. *Census Data API User Guide*. 2019. URL: https://www.census.gov/content/dam/Census/data/developers/api-user-guide/api-guide.pdf (citado na pg. 33).

[Company 2018] The Weather Company. *Weather Company Data for Advanced Analytics*. 2018. URL: https://www.worldcommunitygrid.org/lt/images/climate/The_Weather_Company_APIs.pdf (citado na pg. 37).

[Community 2019] Pandas Community. *Pandas 0.25.1 documentation – Getting started – Package overview*. 2019. URL: https://pandas.pydata.org/pandas-docs/stable/getting_started/overview.html (citado na pg. 11).

[Flater 2011] Drew Flater. *Understanding Geodesic Buffering – Correctly use the Buffer tool in ArcGIS*. 2011. URL: https://www.esri.com/news/arcuser/0111/geodesic.html (citado na pg. 44).

[Google 2019a] Google. *Google Maps Platform: Maps JavaScript API*. 2019. URL: https://developers.google.com/maps/documentation/javascript/places (citado na pg. 39).

[Google 2019b] Google. *Google Maps Platform: Place Types*. 2019. URL: https://developers.google.com/places/web-service/supported_types (citado nas pgs. 39, 41).

[Google 2019c] Google. *Google Maps Platform: Places API*. 2019. URL: https://developers.google.com/maps/documentation/javascript/places (citado na pg. 39).

[Google 2019d] Google. *Google Maps Platform: Places API Policies*. 2019. URL: https://developers.google.com/places/web-service/policies (citado na pg. 39).

[Google 2019e] Google. *Google Maps Platform: Web Services – Elevation API*. 2019. URL: https://developers.google.com/maps/documentation/elevation/start (citado na pg. 41).

[Google 2019f] Google. *Nível gratuito do Google Cloud Platform: Perguntas frequentes*. 2019. URL: https://cloud.google.com/free/docs/frequently-asked-questions (citado na pg. 39).

[Metropolitano de São Paulo 2019] Companhia do Metropolitano de São Paulo. *Pesquisa Origem Destino 2017 50 Anos: A mobilidade urbana da Região Metropolitana de São Paulo em detalhes, Versão 4*. Jul. de 2019. URL: http://www.metro.sp.gov.br/pesquisa-od/arquivos/Ebook%20Pesquisa%20OD%202017_final_240719_versao_4.pdf (citado na pg. 19).

[Sampaio 2018] Cleuton Sampaio. *Data Science para Programadores: Um guia completo utilizando a linguagem Python*. Editora Ciência Moderna, 2018 (citado na pg. 11).

[Yaser S. Abu-Mostafa 2012] Hsuan-Tien Lin Yaser S. Abu-Mostafa Malik Magdom-Ismail. *Learning from data: a short course*. AMLbook.com, 2012 (citado nas pgs. 6, 9, 10).